Bayesian models of learning and generating inflectional morphology

by

Blake H. Allen

A.B., Harvard University, 2011

A.M., Harvard University, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Linguistics)

The University of British Columbia

(Vancouver)

October 2016

Abstract

In many languages of the world, the form of individual words can undergo systematic variation in order to express concepts including tense, gender, and relative social status. Accurate models of these *inflectional* systems, such as verb conjugation and noun declension systems, are indispensable for purposes of both language research and language technology development.

This dissertation presents a theoretical framework for understanding and predicting native speakers' use of their languages' inflectional systems. I propose a probabilistic interpretation of the task that speakers face when inferring unfamiliar inflected forms, and I argue in favor of a Bayesian approach to modeling this task. Specifically, I develop the theory of *sublexical morphology*, which augments the Bayesian approach with intuitive methods for calculating necessary probabilities. Sublexical morphology also possesses the virtue of computational implementability: this dissertation defines all data structures used in sublexical morphology, and it specifies the procedures necessary to use a model for morphological inference. I provide along with this dissertation a Python package that implements all the classes and methods necessary to perform inference with a sublexical morphology model. I also describe an implemented learning algorithm that allows induction of sublexical morphology models from labeled but unparsed training data.

As empirical support for my core claims, I describe the outcomes of two behavioral experiments. Evidence from a test of Icelandic speakers' inflection of novel words demonstrates that speakers are able to additively make use of information from multiple provided inflected forms of a word, and evidence from a similar test on Polish speakers suggests that speakers may be limited to this additive way of combining such pieces of information. In clear support of a Bayesian interpretation of morphological inference, both experiments additionally demonstrate that prior probabilities—understood as reflecting lexical frequencies of different groupings of words—play a major role in speakers' use of their inflectional systems. This is shown to be true even when influence from prior probabilities results in speakers apparently deviating from exceptionless lexical patterns in those systems.

Preface

This dissertation is original intellectual product of the author, Blake Allen. Data about the grammar and lexicon of Icelandic were compiled with assistance from Gunnar Ó. Hansson, who also provided guidance as a native speaker of Icelandic when I was designing the Icelandic experiment. Paulina Lyskawa provided guidance as a native speaker of Polish when I was designing the Polish experiment. The learning algorithm for sublexical phonology grammars, which is the basis of the PyParadigms learning algorithm described in chapter 2, was developed in collaboration with Michael Becker.

Parts of chapters 3 and 4 were presented in their preliminary versions at the International Morphology Meeting (Feb. 2016 in Vienna) and the Germanic Linguistics Annual Conference (May 2016 in Reykjavík).

All projects and associated methods were approved by the University of British Columbia's Research Ethics Board [certificate #H14-01142].

Table of Contents

List of Figures

Glossary

lexeme a unit of meaning with an associated syntactic category, which in languages exhibiting inflection of that syntactic category, specifies a particular phonological form only when associated with a set of morpho-syntactic/semantic features
• Examples: CAT, JUMP

(morpho-syntactic/semantic) feature a category of "auxiliary meanings" which specifies one dimension in an inflectional paradigm
• Examples: *person*, *number*, *tense*, *mood*

(feature) value a meaning specifying one possible semantic referent sub-category for a feature
• Examples: *3rd (person)*, *singular (number)*, *preterite (tense)*, *subjunctive (mood)*

(word) form a phonological shape corresponding to the pairing (combination) of a lexeme and a full set of feature specifications
• Example: in English the plural (number) form of CAT is [kæts]

cell a full set of features which together could specify a form of any lexeme in an inflectional system
• Example: the dative (case) singular (number) cell in the Icelandic noun inflectional system

paradigm the set of all cells in an inflectional system, or a particular lexeme's set of all inflected forms

• Example: in English the paradigm of a noun includes only a singular form and a plural form; the paradigm of CAT is {singular: [kæt], plural: [kæts]}

candidate one of the forms that could conceivably express a particular combination of a lexeme and a set of feature values

• Example: [dʒɪræfs] and [dʒɪrævz] are candidates for the plural form of GIRAFFE

base in the context of a derivation/inference task, a known form of the target lexeme which can be used to infer unknown forms (derivatives) of that lexeme

• Example: a Spanish speaker may use the 1st person singular present and 3rd person singular present forms of the lexeme LOVE, [amo] and [ama] respectively, as bases when attempting to infer that lexeme's 1st person plural present form

• Note: can also be used to refer to a specific cell, abstracted away from any particular lexeme, e.g. "speakers tend to use the infinitive as a base"

derivative in the context of a derivation/inference task, an unknown form of some lexeme which must be inferred

• Example: a Spanish speaker may use known forms of the lexeme LOVE to infer a derivative form of that lexeme such as its 1st person plural present form

• Note: can also be used to refer to a specific cell, abstracted away from any particular lexeme, e.g. "speakers used the singular base forms provided to infer plural derivatives"

constraint a function that assesses whether a form meets some criterion; equivalent to a *feature* (in the Machine Learning sense, not the phonological sense) or an operationalized *independent variable* (in the statistics sense, e.g. as part of a linear model)

• Example: the constraint [3Sg: a#] evaluates to 1 if applied to a 3rd person singular form ending in [a] and evaluates to 0 otherwise

Acknowledgments

Even from before my first day as a PhD student, I have been profoundly fortunate to have the academic and personal support of many of the finest people I have ever known. Thanks in large part to these bonds, I have mercifully been spared most of the hardships usually associated with completing a doctoral degree program.

First, I cannot imagine a department more worthy of being a source of pride for me than UBC's department of linguistics. The hours, thought, and care afforded me by my department-internal dissertation committee members—Gunnar, Kathleen, and especially Doug—have far exceeded even my optimistic expectations from six years ago. It has been a complete pleasure and honor working with all of you as teacher/student and as collaborators, and I sincerely hope that we will be able to continue these relationships even now that I have left Vancouver. At the absolute least, you will always remain, in my thoughts and my heart, the greatest linguists I could have selected to serve on my committee. Speaking of my department, I can honestly say that every professor and staff member there has been a positive influence on me, especially Carla, Bryan, Martina, Eric, Shaine, Strang, and Edna. I am so grateful to all of you and the rest of the department for your role in my life these past five years.

Many mentors outside my department have also played crucial roles in helping me achieve my doctoral degree. Out of all these I first want to thank Michael for helping me take my initial steps toward becoming a fully fledged computational linguist over breakfast at the 2012 LSA meeting in Portland (not to mention his exemplary work as my AM advisor). I am also so fortunate to have spent time discussing my ideas with Alex—the ideal statistician to serve on a linguistics PhD committee—and with Paul, whose sabbatical at UBC was crucial in solidifying

my understanding of the math underlying some of my proposals. I would like to give my thanks to Bruce, Robert, and Naomi for their thoughtful comments and teachings at various points in my program as well, and especially to Adam for the same and for serving as my external examiner. I also acknowledge the friendly help that Paulina provided as I was constructing the methodology for my Polish experiment, and the excellent work of Bosung and the rest of the eNunciate team.

Of course, I would not have made it through my program so happily without the other graduate students who have become such close friends of mine and have shared so many unforgettable hours with me. Zoe, Kamila, Somaye, Erin, Kevin, Andrei, Natalie, Michael (all three of you), and everyone else who has enriched my daily life: thank you! Aside from my fellow UBC graduate students, I cannot go without thanking Sarah for her irreplaceable role in my life: past, future, and *kairos*. And finally, I am tremendously grateful for the unyielding love, enthusiastic support, and exemplary kindness I have received from Ben these past three and a half years.

I will close by noting my greatest appreciation to the two people who have influenced and cared for me the most throughout my entire life: my parents. Mom, Dad, there are no two people in the world whose deep love I have felt so keenly for so long, nor any two who I imagine could have provided as much care and inspiration as the two of you have. Thank you.

Chapter 1

Goals and motivations

Systems of inflectional morphology, including the systems of verb conjugation and noun declension widespread among the world's languages, exemplify the power and complexity of natural language. These systems encode a bidirectional mapping between sound and meaning, and like natural language more generally, they are characterized by their *productivity*: knowledge of an inflectional system gives speakers the ability to generate and comprehend previously unknown sound–meaning mappings.

Unlike the syntax of a language, however, inflectional morphology permits a clearly demarcated, if still infinite, range of meanings. Inflected words typically have a core meaning with an associated syntactic category, such as CAT (noun) or RUN (verb), along with additional elements of meaning which narrow the word's range of denotations, e.g. by specifying a noun as plural in number or a verb as present in tense. In this dissertation, I refer to the core meanings as *lexemes* and the additional elements of meaning as morpho-syntactic/semantic *features*. Each feature can take various *values*; for example, the inflectional system of nouns in English includes a *number* feature which can take two values: *singular* or *plural*.[1] Inflectional systems are systems by which the form (spoken or written) of a lexeme can change depending on its feature values.

The uniquely restricted nature of inflectional systems emerges from a notewor-

[1] The glossary starting on page x contains a full list of related terms and their definitions as used in this dissertation.

thy distinction beween lexemes and features. In any inflectional system, the set of lexemes is *open*, meaning that new lexemes (like TO GOOGLE) can be introduced and assimilated into the inflectional system at any time. Conversely, the set of features and the sets of their respective values are *closed*; except in cases of language change, they cannot be augmented or reduced. Therefore the maximum number of different phonological or orthographical *forms* a lexeme can take is equal to the number of combinations of the feature values its inflectional system includes.[2]

As a result of the semi-closed nature of inflectional morphology, descriptions of inflectional systems often depict them in a tabular format. Each feature corresponds to a dimension (e.g. the horizontal dimension ranging across columns or the vertical dimension ranging across rows), and each value of a feature corresponds to a particular column/row/etc. of the feature's dimension. Figure 1.1 exemplifies such a table for a small subset of the forms of a single verb in normative European Spanish, using phonological representations of segmental information (Alarcos Llorach, 1994). The horizontal dimension displays the possible values for the *person* feature, and the vertical dimension within each sub-table displays the possible values for the *number* feature. The top and bottom sub-tables should be considered a third dimension, showing two possible values *present* and *imperfect* for the *tense* feature.

[*present*]	*1st*	*2nd*	*3rd*
singular	amo	amas	ama
plural	amamos	amais	aman
[*imperfect*]	*1st*	*2nd*	*3rd*
singular	amaba	amabas	amaba
plural	amabamos	amabais	amaban

Figure 1.1: A tabular representation of part of the paradigm for the lexeme TO LOVE in normative European Spanish.

[2]Strictly speaking, the set of phonologically distinct forms can exceed this number in cases of variation, e.g. *smelled* and *smelt* would occupy the same cell in a paradigm. The number of sets of forms with inflectionally distinct meanings, however, is fixed.

If we consider only the feature values shown above, then this table exhausts all of the possible meanings of verbs with the lexeme TO LOVE, as well as all of their canonical forms. It is likely that an adult native speaker of this variety of Spanish will have heard all of these forms at one point or another, and as a result, the ability of such a speaker to produce [amamos] with the intent of conveying the meaning "we love" could be attributed to a feat of memory.

Suppose, however, that a particular native speaker of this variety of Spanish has never encountered the second person plural imperfect of TO LOVE. Because of the productivity of inflectional morphology, this speaker, if pressed by conversational context to produce the appropriate form of this lexeme, would most likely be able to infer from the inflectional system as a whole and from her/his known forms of the lexeme TO LOVE that this unknown form should be [amabais]. Moreover, a third party who has also never heard this form before would likely be able to infer the form's meaning, independent of its context. Given that the set of lexemes is boundless, and that the set of cells grows multiplicatively with the number of features in an inflectional system, the need for both types of inference is likely common in languages with non-trivial inflectional systems and/or frequent coinage of neologisms.

The first of these inferences—inferring and producing an unknown form in an inflectional system—has been called the *paradigm cell filling problem* (Ackerman *et al.*, 2009; Malouf & Ackerman, 2010). This name draws on the tabular metaphor introduced above: each combination of a lexeme and a set of compatible feature values specifies a single *cell* in such a table, and when a speaker does not explicitly know the form that belongs in any one of these cells, she or he must infer a form to *fill* that gap. Here I define a *paradigm* as the set of all the forms associated with a particular lexeme. Inference of an unknown form represents a "problem" in two senses: first, when a speaker needs to produce an inflected form, if a form with that meaning has not been heard before then the speaker cannot pull it from memory and must instead solve the the problem of determining what that form should be; and second, this phenomenon of speakers generating unknown forms presents a problem for linguists because the mechanism by which it occurs is not well understood.

This dissertation develops a formal and computationally implemented model

of the paradigm cell filling problem as it is faced by native speakers of a language with inflectional morphology. The next section describes why a model of the paradigm cell filling problem is both interesting for theoretical linguists and useful for language-related tasks in the real world. Section 1.2 then proposes a set of specific criteria that such a model should meet. The final section provides a summary of the structure of this dissertation, including the next chapter in which I present a framework for modeling the paradigm cell filling problem.

1.1 Why model the paradigm cell filling problem?

Empirically valid formal models of the paradigm cell filling problem can greatly benefit theoretical linguists' understanding of the faculty of language, and computational implementations of these models are useful for a range of natural language processing tasks. This section briefly explains why readers from both fields should be interested in learning about and participating in the development of such models.

1.1.1 Theoretical linguistics

Theoretical linguists rarely describe their research as dedicated to modeling a specific linguistic task or linguistic behavior of speakers; instead, it is common to focus on modeling the faculty of language or grammar itself, with the mostly implicit understanding that an accurate model of the grammatical system itself will naturally explain specific linguistic behaviors related to it. Even so, since at least the rise of generative grammar (Chomsky & Halle, 1968; Chomsky, 1956, 1957), one type of linguistic behavior has claimed central importance as the object of study: a native speaker's inference (or *generation*) of wellformed linguistic units in her or his language. Optimality theory (McCarthy & Prince, 1993; Prince & Smolensky, 2008) and minimalist syntax (Chomsky, 1995), for example, are primarily theories of *derivations* of utterances from the properties of a grammar, and the practical machinery of these theories is crafted for this purpose. However, inference (or generation or derivation) of wellformed utterances is not the only natural language task humans must perform. Such robust, practical machinery for modeling some of these other tasks, such as learning a first language (Pater & Tessier 2003, Hudson Kam & Newport 2005 among others) and judging wellformedness (Coleman & Pierrehumbert 1997, Hayes & Wilson 2008 among others), have been developed to

a degree by communities of linguists, but others, such as identifying the semantics of an encountered sentence/form, have received little attention within theoretical linguistics.

Studying the paradigm cell filling problem is tantamount to studying the familiar topic of language generation or derivation in the domain of inflectional morphology. In this sense, despite the mild unorthodoxy of explicitly describing my research as focusing on a single language-related task, this dissertation has the same goal as most of the existing linguistic theory literature: to explain how speakers use their grammars productively to create wellformed but novel utterances. Making the limited scope of my investigation clear this way therefore does little to actually reduce the breadth of my domain of inquiry, while it does assist in developing a precise theory by creating a laser-like focus on a single linguistic phenomenon.

Moreover, the generation of inflectional morphology in particular warrants study. Inflectional morphology lies at one interface among phonology, semantics, and syntax, and so at the very least, empirical investigations of inflection and a formal understanding of its limitations both serve the purposes of all these sub-fields of linguistics. Largely due to its uniquely semi-closed nature as described above, I agree with the likes of Matthews (1972), Zwicky (1985), Spencer (1991), Anderson (1992), Aronoff (1994), and Beard (1995) that researchers can arrive at useful, illuminating conclusions about language by investigating inflectional morphology on its own rather than only as it relates to phonology or syntax. This dissertation also presents questions about inflectional morphology which research on syntax or phonology would not be likely to ask, and indeed I show that these questions have answers not found in existing literature.

1.1.2 Natural language processing

The paradigm cell filling problem, as I use the term, is equivalent to the task of natural language generation in the domain of inflectional morphology. While the inflectional morphology of English may be simple, the proliferation of computers and internet access across the globe has created a need for language technologies that cover other languages—not only those with large speaker populations, but also those with fewer speakers, e.g. for detecting natural disasters based on social media data (Gales *et al.*, 2014; Ji *et al.*, 2014; Mortensen *et al.*, in review). Most of these

languages are more inflectionally complex than English (Stump & Finkel, 2013).

For these inflectionally complex languages, and even for English, merely compiling a database of inflected forms based on dictionaries or corpora does not suffice for supporting natural language processing systems. There are several reasons that such an approach cannot succeed. The set of lexemes in a language is not fixed, and neologisms that must be integrated into inflectional systems arise frequently, stymying efforts to create a comprehensive database of inflected forms. Moreover, because native speaker use of an inflectional system can differ from prescriptively "correct" forms found in reference texts, and because speaker use of an inflectional system can also change over time, such a database would need to draw on an ever-changing corpus of speaker productions. However, given the roughly Zipfian distribution of lexeme frequencies (Wyllys, 1981), continually expanding such a corpus in order to fill in previously unattested forms is likely to continually add new lexemes—each with, perhaps, only a single form attested—meaning that a corpus-based database of all inflected forms in a language is untenable. In addition, when memory limits are strict, the ability of generative models to produce an infinite number of possible forms based on a constant-sized grammar may be useful. Considering these challenges, computational models of inflectional morphology capable of generating novel forms are essential to natural language generation, especially for inflectionally complex languages.

Many natural language processing tasks including machine translation and question answering require not (only) the generation of inflected forms from a semantic input, but the identification of the semantics of provided inflected forms. Whereas a model that parses the semantics of inflected forms cannot generate novel forms, generative models of the paradigm cell filling problem can, indirectly, serve as models of the opposite task. With a training set of inflected forms and a generative model based on those forms, one can predict all unattested inflected forms for observed lexemes, and these predictions can be matched to inflected forms in some new source text in order to determine the forms' ranges of possible interpretations. In this sense, models of the paradigm cell filling problem can constitute general-purpose models of a variety of inflectional morphology tasks.

1.2 Model desiderata

While models of the paradigm cell filling problem stand to benefit both theoretical linguistics and language technology, these benefits will be maximized only if such models are designed conscientiously, with an eye toward ways they might be used. This section sketches some of the key criteria for a successful model of the paradigm cell filling problem.

1.2.1 Accuracy and precision

It is universally true that useful models must be as accurate as possible, in the sense that for any set of inputs, the outcomes predicted by a model should ideally be identical to the outcomes observed in the system being modeled. I contend that models should also be as *precise* as possible, in the sense that for any set of inputs, the range of outcomes compatible with a model's predictions should be maximally narrow. The importance of this property derives from the fact that more precise models are easier to falsify and therefore easier to improve upon. Even better, a model should make predictions at various levels of precision so that the specific level at which it errs can be identified.

Assessing the accuracy of a model of the paradigm cell filling problem is not as straightforward as ensuring that native speakers produce the same single inflected form as the model predicts for a particular paradigm cell filling problem/query. Speaker behavior in morpho-phonological tasks exhibits widespread but principled variability and gradience (Batchelder 1999; Becker *et al.* 2012; Becker & Gouskova 2013; Hayes & Londe 2006; Hayes *et al.* 2009; Eddington *et al.* 2013; among others). An accurate model therefore cannot predict only a single output or response for each query—it must define a set of possible outputs/responses and a measurement of how likely speakers are to produce each one. Probability theory naturally fits this need: creating probabilistic models allows the assignment of probabilities to particular patterns of outputs given a model, as well as sampling from a model in order to simulate speaker behavior. Given this property of probability theory, as well as the fact that its mathematical bases are well understood and the fact that implementations of various probabilistic model families are widely available, I suggest that the most accurate models of the paradigm cell filling problem must be probabilistic ones.

The accuracy of models of the paradigm cell filling problem should be assessed primarily using behavioral experiments, i.e. *wug* tests (Berko, 1958; Kawahara, 2011, 2016), which elicit judgments about novel inflected forms to determine speaker knowledge of morphological patterns. Ultimately, the system being modeled is the speaker's grammar, that is, her or his ability to use an inflectional system productively, applying it to produce inflected forms that have not previously been encountered. The problem with measuring the accuracy of a model against frequencies (which can be viewed as proportional to probabilities) of forms in a corpus of inflected forms—even in cases where such corpora exist—is that speakers producing inflected forms may simply be retrieving them from memory after having encountered them before. Wug tests provide a convenient and established methodology for avoiding this confound. Wug tests can also be carried out at large scales, for example using internet-based methods, yielding a large sample of responses which can be used to estimate response distributions comparable to the probability distributions predicted by a model.

In the context of the paradigm cell filling problem, a maximally precise model would predict exactly the same inflected form given a grammar and a set of inputs (i.e. base forms of the target lexeme) as speakers with that grammar and those inputs would produce. Given the gradient nature of linguistic behavior, as stated above, the model should actually predict a probability distribution over speaker productions. This level of precision in predictions differs from, for example, a model which simply generates the set of likely inflected forms. Coarser granularities of predictions are still useful, however. As long as a model produces predictions at this maximal level of precision, it can also predict, for example, that some factor x should influence speaker behavior, or that factors y and z should interact in influencing behavior. The nature of these coarser predictions depends largely on the formalism used for a model, but can include, as I address in this dissertation, the model taking a stance on the influence of lexical frequencies on speaker behavior.

1.2.2 Computational implementability

Whereas most theoretical models of inflectional morphology have been defined only in terms of prose descriptions and formal notation, researchers have recently

begun implementing their models computationally, i.e. formalizing them in a body of source code. These computationally implemented models include Network Morphology (Brown & Hippisley, 2012), the generative component of the Minimal Generalization Learner (Albright & Hayes, 2002), and sublexical phonology (Allen & Becker, in review; Gouskova & Newlin-Łukowicz, 2013). No such approach has yet explicitly modeled the paradigm cell filling problem in the domain of entire inflectional systems, but I propose that computational implementability is essential in models of this phenomenon as well.

Creating a computational implementation along with a traditional prose and notational description of a model family confers several benefits. For one, a computational implementation minimizes model ambiguity—other researchers can investigate any aspects of a model by looking at their implementations, which must be clearly defined enough for a computer to run them. The need to achieve this level of clarity also benefits the originator of the theory, sometimes bringing to light ambiguities that would have otherwise gone unnoticed and unaddressed.

Computational implementations also make it far easier to test a model's predictions about a given dataset, even for large datasets, as compared to needing to create a model's representations and work through a model's processes by hand. Consequently, it is possible to test models on much more wide-ranging datasets, both for the researcher developing a model and for others testing it on their own data. (Responses to this need for machine-readable data files also improve the ecosystem of data available to the community of researchers.) Quantitative and probabilistic models in particular virtually require computational implementations, as the price paid for their flexibility and power is a proliferation of mathematical operations necessary when evaluating model predictions.

1.2.3 Learnability

As described above, a precise model of the paradigm cell filling problem will predict a probability distribution over inflected derivative forms given a grammar and a set of base forms. However, quantitative models tend to have very large hypothesis spaces. Unlike an Optimality Theory grammar (McCarthy & Prince, 1993; Prince & Smolensky, 2008), for example, which has $n!$ possible configurations for the $n!$ rankings of n constraints, a grammar with n *weighted* constraints such as

a Maximum Entropy harmonic grammar (Goldwater & Johnson, 2003; Hayes & Wilson, 2008) or other flavor of harmonic grammar (Legendre *et al.*, 1990; Pater, 2009) has an essentially infinite number of possible configurations, since each constraint can take any real number as its weight. One appealing way to cope with this problem of massive hypothesis spaces is to also model the *learning* of models or grammars from sets of input data: defining a procedure for determining a model's parameter values tightly limits the space of valid model configurations for a given set of data. These limitations free the analyst from needing to consider the space of implausible model parameterizations, and can result in more accurate models by including predictors or interactions among predictors that may be difficult for human analysts to notice or formalize (Hayes & Wilson, 2008; Hayes & White, 2013). Moreover, when the input to a model is a set of observables (e.g. the word forms a speaker knows) as opposed to abstract parameters of a grammar, the model can be used more easily for practical purposes, as there is less need for an expert analyst with knowledge of how to tune model parameters.

Beyond these pragmatic reasons, the call for morphological and phonological theory grounded in considerations of learnability by humans is stronger than ever (Albright & Hayes, 2011; Archangeli & Pulleyblank, 2012; Moreton & Pater, 2012) and I find these calls as justified for the paradigm cell filling problem as for other domains of language. One goal of theoretical linguistics is to study the mental linguistic systems of humans, and so since humans learn language from data in our environments, unlearnable grammar formalisms are unlikely to accurately model human language. The domain of learnability has also proved worthy of study in its own right, since some notable aspects of natural language may only become apparent when studied from the standpoint of learnability (Hudson Kam & Newport, 2005; Jesney & Tessier, 2009; McMullin, 2016).

1.3 Structure of dissertation

In the next chapter, I introduce sublexical morphology, a Bayesian framework for modeling the paradigm cell filling problem which is computationally implemented and comes equipped with a learning algorithm. The remainder of the dissertation substantiates the claims that make up the sublexical morphology proposal. Specifically, the content of the following chapters can be summarized as follows.

Chapter 2 lays out my modeling proposals. It starts with a probabilistic interpretation of the paradigm cell filling problem and then introduces a Bayesian view of how speakers "solve" this "problem". Finally and most substantially, the chapter details the sublexical morphology framework which grounds the abstract Bayesian approach in concrete methods for generating inflected forms and inferring their probability distributions.

Chapter 3 presents two experimental investigations into questions of how native speakers use knowledge of base forms when inferring unknown derivative forms, questions whose answers bear directly on the validity of sublexical morphology. First, I use evidence from a behavioral experiment on Icelandic speakers to show that speakers are able to combine discrete pieces of information from multiple inflected base forms when performing inference. Second, I discuss results from a similar experiment on Polish speakers which suggest that speakers may be restricted to linear (additive) combinations of such pieces of information.

Chapter 4 revisits the Icelandic and Polish experimental results, performing post-hoc analyses suggesting that speakers are strongly influenced by raw lexical frequencies of morphological exponents. According to these results, these influences can even cause speakers to fail to productively apply otherwise exceptionless morphological patterns. In the context of sublexical morphology and Bayesian morphology in general, these results support the hypothesis that prior probabilities of morphological exponents play a central role in determining linguistic behavior.

Chapter 5 concludes the dissertation. First it summarizes the claims of previous chapters, interspersing proposals and their empirical support. The chapter then describes how theoretical linguists could fruitfully apply the theory of sublexical morphology to research topics other than the paradigm cell filling problem *per se*, such as paradigm leveling and paradigmatic gaps. It ends by mentioning limitations of the theory as it stands now and proposing follow-up research that could address them and extend the theory's applicability.

Chapter 2

Bayesian morphology and sublexical morphology

This chapter develops the primary claims of the dissertation. First, in section 2.1, I use the language of probability theory to establish a formal description of the paradigm cell filling problem. Section 2.2 then introduces the concept of a surface-oriented, Bayesian view of morphological inference. The most fundamental proposal I make in this dissertation is that such a Bayesian account of the paradigm cell filling problem is both valid and useful, and so this section sets up specific claims that I substantiate in chapters 3 and 4. Because the framework of Bayesian morphology does not by itself constitute a concrete, computationally implementable theory, in the remainder of the chapter, I propose a specific flavor of Bayesian morphology: sublexical morphology. Section 2.3 describes the architecture of sublexical morphology models, and section 2.4 illustrates how such a model can perform the inference necessary to solve the paradigm cell filling problem. The sublexical morphology proposal introduces further empirical claims which are addressed in chapters 3 and 4. Section 2.5 details a learning algorithm for sublexical morphology models. Finally, section 2.6 compares sublexical morphology to a selection of other theories of inflectional morphology.

As a supplement to this dissertation, I have created a Python (van Rossum & Drake, 1995) implementation of sublexical morphology, which is publicly available under the project name *PyParadigms*. This implementation includes a learn-

ing algorithm for sublexical morphology models (described in section 2.5) as well as various command-line utilities for using learned models, e.g. for performing paradigm cell filling problem queries and investigating formal properties of an inflectional system. This software provided freely on an open source basis, available for download at https://github.com/bhallen/pyparadigms.

In this chapter and the remainder of the dissertation, I use a standard set of typographical conventions as laid out here. Names of lexemes like CAT are typeset in small caps, and verb lexeme names are normally written with a preceding "to" as in TO LOVE, but they may sometimes be written simply as LOVE when part of speech is clear from context. Transcriptions are provided in square brackets using IPA symbols, e.g. [kæt], and reflect phonological forms (broad transcriptions) except where otherwise indicated. Orthographical forms like *vivir* are written in italics. Conventions for labeling cells in a paradigm are introduced as necessary throughout, but in general I use compounds of abbreviated names of a cell's morphosyntactic features, e.g. 1SgPresIndic for the first singular present indicative cell.

Throughout this chapter, I will use the system of verbal inflection in normative European Spanish (Alarcos Llorach, 1994) to illustrate key concepts. Specifically, I draw examples from the present indicative cells of the three "regular" classes of verbs: "-ar" verbs (e.g. *amar* TO LOVE), "-er" verbs (e.g. *temer* TO FEAR), and "-ir" verbs (e.g. *partir* TO SPLIT/DEPART). Having limited the domain of examples this way, I use person-number labels like 1Sg and 3Pl to indicate these cells, abstracting away from their tense and mood.

2.1 Formalizing the paradigm cell filling problem

The paradigm cell filling problem is the problem of how to predict the form in an unfamiliar cell of a familiar lexeme's paradigm. I assume that the lexeme must be familiar in the sense that at least one of its forms is known to the speaker or provided to the model.[1] In the context of a probabilistic grammar, this task can

[1]The related question of how a speaker might initially fit some new word into an inflectional paradigm lies outside the scope of this thesis. However, in many cases, this process has been described by other authors. In Spanish, for example, new verbs can enter the inflectional system as infinitives by concatenating the basic pronunciation of the referent with a standard [ear] suffix, e.g. *faxear* TO FAX and *bloguear* TO BLOG (Honrubia *et al.*, 2011). Similarly, in Japanese, a novel verb's dictionary form can be created by replacing the final mora of the base word with the suffix [ru] and enforcing relevant verbal phonotactics, as in [jafuː] YAHOO! → [jafuru] TO SEARCH USING

be re-interpreted as one of first inferring a probability distribution $p(D)$ over the candidates D for this unfamiliar (derivative) form of the familiar lexeme ℓ, and then sampling from that distribution to select a single form to utter.

In order to infer a distribution over derivative form candidates, a speaker must have encountered at least one "base" form of the same lexeme ℓ. These base forms serve two purposes. First, familiarity with at least one such form makes the speaker aware of the lexeme's existence, a logical precursor to solving the paradigm cell filling problem as conceived here. Second, it is the phonological shapes of these base forms that provide information the speaker can then use to infer a specific distribution over derivative forms. For example, a speaker of Spanish who has heard of the lexeme TO BLOG only that its first person singular present indicative form is [blogeo] and that its third person singular present indicative form is [blogea] can infer from the shapes of these forms that the probability of this lexeme's infinitive being [blogear] is much higher than its probability of being [bloger]. Of course, to make productive use of such implicational relationships, the speaker must also have a grammar that encodes them; the general idea of Bayesian morphology does not require any specific formalism for this grammar, and the question of what a grammar compatible with Bayesian morphology might look like is addressed in discussions of sublexical morphology starting in section 2.3.

Just as D represents the set of derivative form candidates d for a lexeme ℓ, I use B to signify the set of possible base forms b in a single cell for a lexeme ℓ. Because forms associated with multiple base cells can conceivably be used in inferring a single derivative form, I employ a subscript to indicate the base form sets of individual base cells. In the abstract, then, we can indicate a speaker's inferred probability that the derivative form of some lexeme ℓ is d, given the observed forms in n base cells, as shown in 2.1.

$$p(D = d|B_1 = b_1, B_2 = b_2, ...B_n = b_n) \tag{2.1}$$

YAHOO! and [gu:guru] GOOGLE → [guguru] TO SEARCH USING GOOGLE (Tsujimura & Davis, 2011).

Hereafter, I will make use of the shorter notational convention shown in 2.2 where contextually appropriate. By this convention, the probability $p(D = d)$ that the discrete variable D takes the value d (e.g. that the plural form of a lexeme has some particular phonological form) is abbreviated as $p(d)$, with equivalent abbreviations for other variables B_1 etc.

$$p(d|b_1, b_2, ...b_n) = p(D = d|B_1 = b_1, B_2 = b_2, ...B_n = b_n) \tag{2.2}$$

For example, either of these formats can represent the Spanish TO BLOG example introduced above, as shown in equation 2.3.

$$\begin{aligned} & p([blogear]|[blogeo], [blogea]) \\ = p(infinitive = [blogear]|1Sg = [blogeo], 3Sg = [blogea]) & \end{aligned} \tag{2.3}$$

In reality, there is no restriction to a single phonological form for any lexeme-cell combination. There may be multiple phonologically distinct forms in common use, for example [kæktəsɪz] and [kæktaɪ] for the plural of CACTUS in English.[2] The probabilistic framework I have laid out is compatible with this complexity: instead of conditioning derivative candidate distributions on base form variables which must each take a single value, as in 1Sg=[blogeo], we can condition them on observation counts of base forms. Section 2.3 through 2.5 provide more detail about how such counts are used in sublexical morphology. The notation shown in equation 2.3 is more compact and approximates the relevant facts for forms with no phonological variation, and so I continue to use it for expositional purposes. Wherever the simpler notation is used, it can be treated as a shorthand for relevant distributions of frequency counts.

[2] In this dissertation I abstract away from predictable allophonic detail in inflected forms, for example the presence or absence of aspiration on the initial [k] of [kæktaɪ]. I assume that all mental calculations are carried out on phonological forms that have been accurately inferred from interlocutor speech.

2.2 Bayesian morphology

The previous section defined the task of generating an unfamiliar derivative form—that is, solving an instance of the paradigm cell filling problem—as the inference of a conditional probability distribution $p(D|B_1, B_2, ...B_n)$ followed by sampling from that distribution. This definition alone, however, falls far short of a useful model of the phenomenon, as there is no clear way to directly infer such a distribution from base forms. Moreover, to directly calculate derivative probabilities this way would require summing over derivative probabilities conditioned on all possible combinations of base forms in order to arrive at the constant used to normalize probabilities so that they sum to 1. Because any concatenation of phonological units is a possible form of a given lexeme in any specific cell, this set of combinations of possible base forms is infinite, and summing over it would require commensurate computational resources (although see Hayes & Wilson 2008 for an approach that approximates this normalization constant at substantial but finite computational expense). If learning such a model from a set of training data based on counts of base form combinations, sparsity would also pose a substantial problem, as only a tiny fraction of possible combinations of base forms would be likely to be represented in the data.

Fortunately, by applying Bayes's theorem, it is possible to decompose the conditional distribution $p(D|B_1, B_2, ...B_n)$ into sub-parts more amenable to direct calculation. As shown in equation 2.4, the probability of a derivative candidate given a set of base forms is proportional to the following quantity: the probability of those base forms given the derivative candidate, times the prior probability of the derivative candidate.

$$p(D|B_1, B_2, ...B_n) \propto p(B_1, B_2, ...B_n|D)p(D) \tag{2.4}$$

Proportionality of the right-hand side to the left-hand side in this case means that the left-hand side is equal to the right-hand side divided by a normalization constant Z. This constant sums the right-hand side of equation 2.4 across all possible derivative candidates. The set of derivative forms to sum over is still infinite like the normalization constant that would be required to calculate $p(D|B_1, B_2, ...B_n)$. However, given a finite approximation of the set of all possible forms in an arbi-

trary cell, the size of this set for a single derivative cell is smaller than the size of the set of all combinations of base cells by a factor of the number of cells in an inflectional system. More importantly for the purposes of this dissertation, one can posit constraints that restrict the set of derivative candidates in particular to a small set, such that summing over them is trivial; this is the approach that I take in the following section.

I define *Bayesian morphology* as the proposal that the behavior of native speakers faced with the paradigm cell filling problem can be predicted using this application of Bayes's theorem to the probability theoretic definition of the paradigm cell filling problem. This general framework predicts, for example, that knowledge of the prior probabilities of derivative forms (as defined in some way) is indispensable in predicting morphological behavior. But while Bayesian morphology has the computational and empirical advantages described here, it still lacks the specificity necessary to constitute a practical theory of the paradigm cell filling problem. In order to achieve this research goal, it is necessary to create a theory that builds on Bayesian morphology by adding mechanisms for calculating the distributions $p(B_1, B_2, ...B_n|D)$ and $p(D)$.

2.3 Sublexical morphology

Sublexical morphology is a specific flavor of Bayesian morphology that provides an intuitive, computationally implementable, and efficient way to calculate the distributions $p(B_1, B_2, ...B_n|D)$ and $p(D)$ and therefore the paradigm cell filling problem objective distribution $p(D|B_1, B_2, ...B_n)$. The sublexical morphology framework also provides an algorithm for learning models of inflectional systems from sets of *training data*, form–meaning pairs which approximate the information that human learners could plausibly have when learning their morphological grammars. Sublexical morphology is based in part on sublexical phonology (Allen & Becker, in review; Becker & Gouskova, 2013; Gouskova & Newlin-Łukowicz, 2013), from which it inherits the spirit of concepts like sublexicons and gatekeeper grammars, although sublexical phonology lacks the explicitly Bayesian character of sublexical morphology which is a major focus of this dissertation.

This section introduces the central claims specific to sublexical morphology, explaining how they build on the general claims of Bayesian morphology. This ex-

position includes both an introduction to the core theoretical claims of the framework and the data structures that the theory of sublexical morphology uses to represent an inflectional system. Section 2.4 then details the mechanism by which a sublexical morphology model can be used to perform morphological inference, that is, how it can solve the paradigm cell filling problem. Section 2.5 follows up on these descriptions of sublexical morphology models by explaining the algorithm by which such models can be learned from a set of training data. Finally, section 2.6 compares sublexical morphology to some other influential theories of inflectional morphology.

2.3.1 Theoretical core of sublexical morphology

The theory of sublexical morphology can be characterized by the claim that an inflectional system is comprised of a set of *paradigm sublexicons*, each of which is a set of lexemes with identical morphological behavior. These paradigm sublexicons resemble the traditional concept of inflectional classes, but have more clearly defined internal structures and roles in derivation. Cases in which a lexeme exhibits multiple attested forms in any particular cell are the principled exception to this rule, and such a lexeme may belong to multiple paradigm sublexicons; such cases are discussed further later in this section. In general, however, a language's paradigm sublexicons can be thought of as a partition of the lexicon into by-lexeme (*not* by-cell) subparts each of which is homogeneous with respect to the language's inflectional morphology. Note that except where contrasting paradigm sublexicons with the related concept of *mapping sublexicons* in section 2.5, I will refer to paradigm sublexicons simply as *sublexicons*.

When I describe a sublexicon as homogeneous in its morphological behavior, I mean that for every lexeme in a sublexicon, for each pair of cells, there is a single *morphological operation* (which may include multiple changes, e.g. stem vowel mutation and suffixation) that takes as input the phonological form of that lexeme in one of those cells and outputs the phonological form of that lexeme in the other cell. These operations deal only in surface-level phonological forms, not abstract underlying representations or roots; for a discussion of how this approach contrasts with other theories of inflectional morphology and how I justify it, see section 2.6. For example, in normative European Spanish, there might be one sublexicon (the

"-ar" sublexicon) with morphological operations like those shown in 2.5. The # symbol here indicates the right edge of an inflected form.

$$\text{morphological operations}: \begin{cases} \text{1Sg} \rightarrow \text{2Sg}: & [\text{o\#}] \rightarrow [\text{as\#}] \\ \text{1Sg} \rightarrow \text{3Sg}: & [\text{o\#}] \rightarrow [\text{a\#}] \\ \ldots & \\ \text{3Pl} \rightarrow \text{1Pl}: & [\text{n\#}] \rightarrow [\text{mos\#}] \\ \text{3Pl} \rightarrow \text{2Pl}: & [\text{n\#}] \rightarrow [\text{is\#}] \end{cases} \tag{2.5}$$

The central reason for positing sublexicons is that when at least one base form of a lexeme is known, division of the lexicon into sublexicons sets up a direct mapping from sublexicon to derivative candidate, and this mapping can be used in morphological inference. Sublexical morphology allows many-to-one mappings from sublexicons to derivative candidates, i.e. different sublexicons that happen to generate the same derivative candidate, but it explicitly disallows one-to-many mappings, meaning that the choice of a sublexicon fully determines the choice of a derivative candidate. This property also extends to a probabilistic setting: establishment of a probability distribution over sublexicons fully determines a probability distribution over derivative candidates. In general, the probability of a derivative form is equal to the probability of the sublexicon that generates that derivative form, as shown in equation 2.6. This equation constitutes perhaps the most central proposal of this dissertation. Note that there is a trivial exception to this simple equality when multiple sublexicons generate the same derivative form; since their probabilities only need to be added together, for now I abstract away from these edge cases to avoid baroque notational conventions.

$$p(D|B_1, B_2, ...B_n) = p(S|B_1, B_2, ...B_n) \tag{2.6}$$

This equality reduces the task of inferring a probability distribution over derivative candidates to a classification task: the speaker or model needs only to assess how similar the target lexeme (the one whose derivative form is being inferred) is to each sublexicon, and this probabilistic classification suffices to arrive at a distribution over derivative candidates. However, because the distribution over sublexicons

is still conditioned on the joint distribution over all sets of base forms, an infinite set of sets, direct calculation of conditional distributions over sublexicons is still infeasible for the same reason as discussed in section 2.2.

Fortunately, the equality set up between derivative candidates and sublexicons leaves the distribution over the latter equally amenable to application of Bayes's theorem. By applying the theorem just as shown in section 2.2, but with distributions S over sublexicons substituted for distributions D over derivative forms, we arrive at equation 2.7.

$$p(S|B_1, B_2, ...B_n) \propto p(B_1, B_2, ...B_n|S)p(S) \tag{2.7}$$

The advantage of this interpretation of the paradigm cell filling problem is that both quantities from which derivative probabilities emerge, $p(B_1, B_2, ...B_n|S)$ and $p(S)$, have intuitive, computationally tractable methods of calculation. Finally, therefore, this equation is the end result of all the manipulations necessary to describe sublexical morphology, since it achieves the goal of relating easily computable quantities to the quantities of central import, i.e. the probabilities of derivative candidates. The term $p(B_1, B_2, ...B_n|S)$, which I call the *likelihood term* because it indicates likelihood of the attested base forms given a particular sublexicon, can be calculated by using sublexicons' *gatekeeper grammars*, log-linear models based on phonological constraints. The *prior* probabilities of sublexicons $p(S)$ correspond to the "sizes" of the various sublexicons in terms of how many lexemes are associated with them.

The following subsection takes these theoretical claims and specifies how they are implemented in a set of formal, pseudo-computational data structures. This half of the section largely serves to provide a concrete grounding to the abstract claims set up so far, as well as to detail how various special cases are handled. Note that chapters 3 and 4 provide empirical evidence for the likelihood term and the prior term, respectively.

2.3.2 Data structures of sublexical morphology

Within the sublexical morphology framework, a model of a particular inflectional system consists of a set of (paradigm) sublexicons. Each sublexicon has three components: a set of associated lexemes (or, more properly, associated forms of

lexemes), a set of morphological operations that map from forms in one cell to forms in another, and a gatekeeper grammar that assesses the likelihood of a set of base forms given the sublexicon. The structure of a model, and the structures of its sublexicons, are schematized in 2.8. In this subsection I describe each of these parts in turn.

$$\text{model}: \begin{cases} \text{paradigm sublexicon}: \begin{cases} \text{associated forms} \\ \text{morphological operations} \\ \text{gatekeeper grammar} \end{cases} \\ \text{paradigm sublexicon}: \begin{cases} \text{associated forms} \\ \text{morphological operations} \\ \text{gatekeeper grammar} \end{cases} \\ \ldots \end{cases} \tag{2.8}$$

Associated forms

Each paradigm sublexicon is associated with a morphologically homogeneous subset of the lexemes in the language's inflectional system. In the case of the normative European Spanish verbs example—ignoring for now complicating phenomena like diphthongization and velar insertion (Albright, 2002) which would increase the number of sublexicons—the inflectional system can be split into three sublexicons that parallel the three canonical inflectional classes of Spanish verbs:

$$\text{model}: \begin{cases} \text{paradigm sublexicon 1 (“-ar”)} \\ \text{paradigm sublexicon 2 (“-er”)} \\ \text{paradigm sublexicon 3 (“-ir”)} \end{cases} \tag{2.9}$$

Each form among the associated forms of a sublexicon is stored along with a label for its lexeme, which cell it belongs to, and its frequency, as shown in the abstract in 2.10 and for a hypothetical Spanish “-ar” sublexicon in 2.11. Interpreting sublexical morphology as a theory of human use of inflectional morphology, these frequency counts indicate the number of times each phonological form has been heard by a speaker; when a sublexical morphology model is learned by a computational implementation of its learning algorithm, frequencies represent the frequencies of those forms in the training data provided to the learning algorithm.

Either way, I assume in this chapter that forms are stored as phonological representations of uttered inflected forms. Note however that sublexical morphology is not strictly limited to the domain of phonological representations, and can operate over orthographical representations where they approximate phonological representations, as used in chapters 3 and 4.

$$\text{associated forms} : \begin{cases} \text{cell, lexeme, form:} & \text{frequency} \\ \text{cell, lexeme, form:} & \text{frequency} \\ \multicolumn{2}{c}{\dots} \end{cases} \tag{2.10}$$

$$\text{associated forms} : \begin{cases} \text{1Sg, SPEAK, [ablo]:} & 800 \\ \text{2Sg, SPEAK, [ablas]:} & 200 \\ \multicolumn{2}{c}{\dots} \\ \text{2Pl, COOK, [kosinais]:} & 100 \\ \text{3Pl, COOK, [kosinan]:} & 700 \end{cases} \tag{2.11}$$

When each lexeme–cell combination in a language has only one attested phonological form, all forms of a particular lexeme will be associated with just a single sublexicon. Section 2.5, which describes the learning algorithm for sublexical morphology models, makes it clear why this is the case. When there is any variability in the form of a particular lexeme in any cell, however, that lexeme will be associated with one sublexicon for each of its variants. For example, the English plural of CACTUS varies between [kæktəsɪz] and [kæktaɪ], and so the former form would be associated with an "add final -ɪz to pluralize" sublexicon and the latter with an "[əs#]→[aɪ#] to pluralize" sublexicon. It is convenient, though, to think of sublexicons roughly as partitions of the lexicon by lexeme, since describing them as partitions "by form" is less clear about the criteria for the partitioning (and would, e.g. be consistent with mistakenly thinking of each sublexicon as containing a subset of the *cells* in an inflectional system). Note as well that because sublexical morphology lacks any mechanism for performing derivational morphology, even a compound which one could argue "contains" multiple lexemes, like English BLACKBOARD or SUNLIGHT, is itself treated as a single lexeme.

Morphological operations

The forms within a particular sublexicon are morphologically homogeneous with respect to each other, and this property of homogeneity is encoded in a sublexicon's morphological operations. A single morphological operation is defined as the changes that must be applied to the inflected form in one cell to produce its inflected form in a different cell. Therefore there are as many morphological operations associated with a paradigm sublexicon as there are ordered pairs of cells in the inflectional system, a total of $n^2 - n$ operations for n cells. The schematic in 2.12 shows the abstract form of a sublexicon's morphological operations, and 2.13 repeats the earlier example of morphological operations for the "-ar" sublexicon in normative European Spanish. As before, the # symbol is used to indicate the right edge of an inflected form.

$$\text{morphological operations}: \begin{cases} \text{base cell} \rightarrow \text{derivative cell}: & \text{operation} \\ \text{base cell} \rightarrow \text{derivative cell}: & \text{operation} \\ & \dots \end{cases} \tag{2.12}$$

$$\text{morphological operations}: \begin{cases} \text{1Sg} \rightarrow \text{2Sg}: & [\text{o\#}] \rightarrow [\text{as\#}] \\ \text{1Sg} \rightarrow \text{3Sg}: & [\text{o\#}] \rightarrow [\text{a\#}] \\ \dots & \\ \text{3Pl} \rightarrow \text{1Pl}: & [\text{n\#}] \rightarrow [\text{mos\#}] \\ \text{3Pl} \rightarrow \text{2Pl}: & [\text{n\#}] \rightarrow [\text{is\#}] \end{cases} \tag{2.13}$$

The morphological homogeneity of a sublexicon can be formalized in terms of such operations. Each operation can be thought of as a function that takes an inflected base form as its input and yields an inflected derivative form as its output. For a sublexicon to be morphologically homogeneous, then, the following must hold: for any lexeme in that sublexicon, for any base cell and derivative cell, the sublexicon's morphological operation from that base cell to that derivative cell must generate (one of) the lexeme's attested form(s) in the derivative cell when provided (one of) the lexeme's attested base form(s).[3] For example, the operations

[3]The parenthetical additions here accommodate the fact that a lexeme may be associated with

for the Spanish sublexicon shown in 2.13 are valid for all pairs of forms that would be associated with that sublexicon, e.g. TO TOUCH 1Sg [toko] ∼ 2Sg [tokas] and TO SPEAK 3Pl [ablan] ∼ 2Pl [ablais].

Each morphological operation can include multiple individual changes. In a model of the inflectional morphology of Arabic nouns, there might be a sublexicon whose morphological operations include one indexed to a cell *Singular* as the base cell and *Plural* as the derivative cell and comprising the following changes: mutate the first vowel to [a], insert [aː] after the second consonant, and mutate the last vowel to [i]. Such an operation would map between inflected form pairs like LOCUST Singular [dʒundub] ∼ Plural [dʒanaːdib] (Childs, 2003).

Sublexical morphology does not commit to any one particular computational implementation of the functions constituting these morphological operations. They could conceivably be implemented as finite state transducers (Karttunen & Beesley, 2005), at least for regular morpho-phonological relations. The PyParadigms implementation of sublexical morphology follows the work of Allen & Becker (in review) in using an operation formalism designed specifically for encoding morpho-phonological changes in a way that mirrors the cross-linguistic typology of such changes. This formalism makes it possible for operation positions to be stated, for example, in terms of a word's final syllable nucleus. While the formalism used for these operations plays an undeniably large role in the utility and learnability of sublexical morphology models, the task of making empirical claims about the nature of these operations beyond their basic nature as mappings from inflected form to inflected form lies outside the scope of this dissertation. Allen & Becker (in review) includes a discussion of considerations related to this issue.

Gatekeeper grammar

Finally, each sublexicon includes a *gatekeeper grammar*. This component of a sublexicon assigns a probability to a set of provided base forms; intuitively, this probability indicates how well-formed the base forms are *as members of that sublexicon*. Gatekeeper grammars have the formal structure of Maximum Entropy (MaxEnt) harmonic grammars (Goldwater & Johnson, 2003; Hayes & Wilson, 2008; Wilson,

multiple sublexicons. In such cases, for each sublexicon that a lexeme is associated with, there must be some pair of a base form and a derivative form such that the appropriate morphological operation in that sublexicon correctly generates the derivative form from the base form.

2006), meaning that they are constraint-based grammars in which constraints have real-valued weights rather than a set ranking. Outside the domain of theoretical linguistics, they can be described as log-linear models (Knoke & Burke, 1980). A gatekeeper grammar is fully parameterized by its set of constraints and their weights.

Notably, while I borrow use of the term *constraint* from the MaxEnt harmonic grammar literature, these objects are not constraints in the usual phonological sense of the word. Each constraint acts as an *indicator function*, meaning that it evaluates to 1 if a particular structure is present in the input and otherwise evaluates to 0. In the terminology of Optimality Theory (Prince & Smolensky, 2008), evaluating to 1 equates to assigning a violation. Notably, however, a constraint may either serve to detect a structure whose presence *increases* the probability of the input base forms (if the constraint has a positive weight) or to detect a structure whose presence *decreases* that probability (if it has a negative weight). It is also possible to restrict grammars to using only positive or negative weights, or only weights within arbitrary ranges, as desired—see Pater (2009) and Daland (2015) for discussions of some implications of various weight conventions.

Evaluation of a constraint can be usefully thought of as a two-step process, one reflected in the notational convention I use for constraints and in the data structure used to represent them. A constraint first needs a label for the cell whose forms it evaluates; having extracted the input form in that cell (if one is provided), the remainder of the constraint provides a description of the structure whose presence in that form will result in the constraint evaluating to 1. There are no restrictions on which cells a sublexicon's constraints can refer to. As with the morphological operations, sublexical phonology is agnostic as to the specific details of how this structure-detecting component is implemented. PyParadigms uses regular expressions supplemented with rules capable of expanding featural descriptions (if compatible with a provided phonological feature system) into sets of characters amenable to regular expression matching. The abstract template for a gatekeeper grammar and an example of the grammar for the Spanish "-ar" sublexicon are shown in 2.14 and 2.15, respectively. Again, the symbol # indicates the right edge of a word, equivalent to $ in a regular expression.

$$\text{gatekeeper grammar}: \begin{cases} \text{cell: target sequence} & \textit{weight} \\ \text{cell: target sequence} & \textit{weight} \\ & \dots \end{cases} \tag{2.14}$$

$$\text{gatekeeper grammar}: \begin{cases} \text{3Sg: a\#} & 4 \\ \text{3Sg: e\#} & -2.5 \\ \text{3Pl: an\#} & 2.2 \\ & \dots \end{cases} \tag{2.15}$$

The example below provides a tabular representation of how the Spanish "-ar" sublexicon's grammar operates on two Spanish forms. This tableau represents a hypothetical case of the paradigm cell filling problem in which two forms of the verb TO LOVE, 3Sg [ama] and 3Pl [aman], have been provided to the gatekeeper grammar, as one step in the process of establishing a probability distribution over candidates for this lexeme's form in some other cell. Columns in the central part of the table correspond to constraints, whose weights (w) are given in the uppermost row. Each of the two rows below the horizontal rule indicates how a single inflected form has been evaluated by constraints in the grammar. The harmony scores $\mathscr{H}$ of each base form, as well as their total harmony, are shown in the rightmost column. Harmony scores are weighted sums of forms' violation profiles.

		w=4	w=-2.5	w=2.2	w=-1	
	freq.	3Sg: a#	3Sg: e#	3Pl: an#	3Pl: en#	Form $\mathscr{H}$
3Sg: ama	1	1	0	0	0	4
3Pl: aman	1	0	0	1	0	2.2
					Total harmony of bases:	6.2

Figure 2.1: Tableau showing example weights and violation profiles for four hypothetical gatekeeper grammar constraints, as well as the forms' harmony scores. In this example, two bases for a verb have been provided to this grammar as inputs, and the grammar has produced a total harmony score of 6.2 for them.

Because this section focuses only on describing the formal structure of sublexicons, the procedure by which gatekeeper grammars' harmony scores are used to evaluate wellformedness is covered in detail in the section dedicated to inference in sublexical morphology, section 2.4, and the learning algorithm for these grammars will be given a similar treatment in 2.5 along with the other aspects of the overall sublexical morphology learning algorithm.

As a final piece of exposition about the structure of gatekeeper grammars, I note that there is currently no way to explicitly encode dependencies between sublexicons or between their gatekeeper grammars. If the associated forms for two sublexicons have very similar evaluation profiles for the constraints used in their gatekeeper grammars, then the weights of those constraints will be similar in those two sublexicons' grammars. These similarities are therefore not "accidental" because they are derived from and grounded in the phonological similarities of the different sublexicons, but there is no mechanism for grammars to share weights with each other or otherwise directly interact.

2.4 Derivative inference in sublexical morphology

I turn now from a description of the formal structure of sublexical morphology models to a description of the process by which such a model can solve the paradigm cell filling problem. I call this process *derivative inference*, since it results in a conditional probability distribution over candidates for the form of the specified lexeme in the specified derivative cell. Recall that sublexical morphology arrives at this distribution via the the equalities shown in equation 2.16, which—for reasons that will soon become clear—explicitly includes division by the normalization constant Z instead of leaving this division implicit by using the proportionality symbol $\propto$ as shown earlier in e.g. equation 2.7.

$$\begin{aligned} & p(D|B_1, B_2, ...B_n) \\ &= p(S|B_1, B_2, ...B_n) \\ &= \frac{p(B_1, B_2, ...B_n|S)p(S)}{Z} \end{aligned} \tag{2.16}$$

To summarize this equation, the probability of a derivative candidate d given the base forms b_1 through b_n can be calculated from the following quantities:

1. $p(b_1, b_2, ...b_n|s)$, the likelihood of b_1 through b_n given the sublexicon s which generates d

2. $p(s)$, the prior probability of the aforementioned sublexicon s

3. the normalization constant Z

The subsections of this section walk through these three elements of the equation part by part, using an example paradigm cell filling problem query based on Spanish to illustrate relevant mechanics. An additional subsection then addresses the topic of how derivative candidates are generated in sublexical morphology, showing also how a distribution over sublexicons is then used to arrive at a distribution over derivative forms. The final subsection shows how these values are brought together by equation 2.16 for the Spanish example.

2.4.1 Calculating probabilities of bases

The term $p(b_1, b_2, ...b_n|s)$ indicates the *likelihood* (probability) of the observed distributions over base forms *given* (assuming) that the lexeme in question is a member of paradigm sublexicon s. This likelihood can be usefully thought of as a kind of comparative phonotactic wellformedness rating (cf. Hayes To appear) which expresses how much the phonological shapes of the observed bases match the phonological regularities among forms associated with a sublexicon. In sublexical morphology, these likelihood values are calculated by sublexicons' gatekeeper grammars: the joint probability of the observed bases of the target lexeme given a particular paradigm sublexicon s out of the set of paradigm sublexicons S is determined by applying the gatekeeper grammar of s to the observed bases.

In order to unpack this description, I begin by setting up a hypothetical paradigm cell filling problem task (for a speaker) or query (to a model), focusing again on normative European Spanish. Suppose that there exists a speaker of this language who, despite otherwise normal fluency in the language, has limited experience with the lexeme TO LOVE. Particularly, this speaker has only ever heard the infinitive (Inf) of this lexeme, [amar], and its third person singular (3Sg) form, [ama], and each only once. In some conversational setting, this speaker finds a need to express the lexeme TO LOVE in its first person plural (1Pl) form. In other words, the speaker needs to use her/his knowledge of the inflectional system, [amar], and

[ama] in order to infer a probability distribution over the candidates for the 1Pl form of TO LOVE from which she/he can sample part of an utterance. Alternatively, a sublexical morphology implementation of the Spanish verbal inflection system in a computer could be provided the following query: what is the distribution over 1Pl forms for a verb given that its Inf form is [amar] and its 3Sg form is [ama]?

Having established the nature of the example paradigm cell filling problem task/query, consider the three hypothetical sublexicons of regular verbs in normative European Spanish, repeated below as 2.17.

$$\text{model} : \begin{cases} \text{paradigm sublexicon 1 (“-ar”)} \\ \text{paradigm sublexicon 2 (“-er”)} \\ \text{paradigm sublexicon 3 (“-ir”)} \end{cases} \tag{2.17}$$

Each of these sublexicons has its own gatekeeper grammar, the constraint weights of which reflect the phonological regularities of each sublexicon. For example, among lexemes in the “-ar” sublexicon, infinitive forms always end in [-ar], while infinitives in the “-er” and “-ir” sublexicons invariably end in [-er] and [-ir], respectively. 3Sg forms in the “-ar” sublexicon end in [-a], while those in the other two sublexicons end in [-e]. Such categorical regularities, in addition to numerous gradient patterns that distinguish sublexicons, are encoded in their weights. Example weights in the three sublexicons of constraints relevant to these patterns are shown in 2.2.

	Inf: ar#	Inf: er#	Inf: ir#	3Sg: a#	3Sg: e#
-ar sublexicon	**2.3**	-1.1	-1.4	**1.2**	-0.7
-er sublexicon	-0.3	**2.1**	-1.3	-0.8	**1.3**
-ir sublexicon	-0.9	-1.1	**2.4**	-1.0	**1.1**

Figure 2.2: Arbitrary example weights of various constraints in the “-ar”, “-er”, and “-ir” sublexicons of normative European Spanish. *Inf* is used as an abbreviation for the infinitive cell of the paradigm. Positive constraint weights are bolded for ease of visual parsing.

At this point, inference proceeds by having each sublexicon assign a likelihood

to the provided base forms Inf: [amar] and 3Sg: [ama]. Strictly speaking, the likelihood of a set of bases given a sublexicon s is defined as its likelihood given that sublexicon's gatekeeper grammar g:

$$p(B_1, B_2, ...B_n|s) = p(B_1, B_2, ...B_n|g) \tag{2.18}$$

Because all gatekeeper grammars g are Maximum Entropy (MaxEnt) harmonic grammars, the method of assessing probabilities conditional on them is well defined. According to the definition of a MaxEnt harmonic grammar (Goldwater & Johnson, 2003; Hayes & Wilson, 2008; Wilson, 2006), the probability of a phonological object x conditional on a grammar g is equal to the constant e to the power of that object's *harmony score* $\mathscr{H}_{x,g}$, divided by a normalization constant indexed to that grammar Z_g, as shown in equation 2.19.

$$p(x|g) = \frac{e^{\mathscr{H}_{x,g}}}{Z_g} \tag{2.19}$$

In the case of a gatekeeper grammar, the phonological object being evaluated is the set of all n provided base forms $b_1, b_2, ...b_n$. Their cumulative harmony score is equal to the sum of their individual harmony scores. The harmony score of a single base form is equal to the weighted sum of its constraint output profile, i.e. its constraint "violations". Equation 2.20 presents this definition by using $c_{g,i}$ to refer to the output (1 or 0) of the i^{th} constraint in the grammar g and using w to refer to the weight of the i^{th} constraint. b_j indicates the base form in the j^{th} base cell.

$$\mathscr{H}_{b_1, b_2, ...b_{n,g}} = \sum_i \sum_j w_{g,i} c_{g,i} b_j \tag{2.20}$$

In equation 2.19, Z_g is the sum of $e^{\mathscr{H}}$ over all possible forms in the inflectional system. Because of the difficulties inherent in estimating the phonological properties of this infinite set of forms, sublexical morphology follows the example set by sublexical phonology (Allen & Becker, in review) by approximating Z_g as the sum of $e^{\mathscr{H}}$ over all forms *available to the model*.[4] Concretely, then, the value of the

[4]See, however, Hayes & Wilson (2008) for a way to more accurately estimate the phonological properties of an infinite space of forms.

normalization constant Z_g is given by the equality in 2.21, where b ranges across all the base forms in a sublexicon s or in the data provided as part of the paradigm cell filling problem task/query and $\mathscr{H}_{b,g}$ is the harmony of base b given the grammar g of s.

$$Z_g = \sum_s \sum_b \mathscr{H}_{b,g} \tag{2.21}$$

Intuitively, the normalization constant Z_g serves to ensure that the probabilities assigned by g to the set of possible base forms constitute a proper probability distribution by summing to 1.0. This purpose is similar to that of the distinct normalization constant Z described in subsection 2.4.3, which normalizes probabilities of paradigm sublexicons rather than those of base forms.

Returning to the Spanish example, in order to determine the likelihood of the three sublexicons in 2.17, each sublexicon must assign a probability to the provided forms Inf: [amar] and 3Sg: [ama]. Figure 2.3 shows, for each sublexicon, the constraint output/violation profiles for these forms, as well as their individual and cumulative harmony scores and their overall likelihood. For the weights of these constraints in the different sublexicons, see Figure 2.2.

-ar sublexicon

	Inf: ar#	Inf: er#	Inf: ir#	3Sg: a#	3Sg: e#	$\mathscr{H}$
Inf: [amar]	1	0	0	0	0	2.3
3Sg: [ama]	0	0	0	1	0	1.2
					Cumulative harmony:	3.5
					Likelihood:	0.331

-er sublexicon

	Inf: ar#	Inf: er#	Inf: ir#	3Sg: a#	3Sg: e#	$\mathscr{H}$
Inf: [amar]	1	0	0	0	0	-0.3
3Sg: [ama]	0	0	0	1	0	-0.8
					Cumulative harmony:	-1.1
					Likelihood:	0.003

-ir sublexicon

	Inf: ar#	Inf: er#	Inf: ir#	3Sg: a#	3Sg: e#	$\mathscr{H}$
Inf: [amar]	1	0	0	0	0	-0.9
3Sg: [ama]	0	0	0	1	0	-1.0
					Cumulative harmony:	-1.9
					Likelihood:	0.001

Figure 2.3: Constraint output/violation profiles for the input comprising Inf: [amar] and 3Sg: [ama], for the three example sublexicons. Constraint weights are given in figure 2.2. It is assumed that all three sublexicons have a normalization constant Z_g of 100, although this constant will normally vary from sublexicon to sublexicon.

Once the likelihood of the provided base forms given each sublexicon has been calculated, the role of the gatekeeper grammars has ended. These values are stored until the sublexicons' prior probabilities and their normalization constants are calculated, if they have not been calculated already, so that these values can be combined afterwards.

2.4.2 Calculating prior probabilities

In sublexical morphology, calculating the prior probability of a particular paradigm sublexicon is intuitive and computationally simple. This probability corresponds to

the "relative size" of a paradigm sublexicon, which is a function of the frequencies of the forms in each paradigm sublexicon. In the PyParadigms implementation, the relevant frequencies are assumed to be type rather than token frequencies, i.e. the cardinality of the set of a paradigm's associated forms.

As an illustration of this principle, 2.22 shows an example of the forms that could be associated with the hypothetical "-ar" sublexicon, repeated with slight modification from 2.11. The overall type frequency of this sublexicon is 4, since it contains a total of four forms. If the "-er" and "-ir" sublexicons contained 5 and 7 distinct forms, respectively, then the prior probability of the "-ar" sublexicon would be $4/(4+5+7) = 0.25$.

$$\text{associated forms} : \begin{cases} \text{1Sg, SPEAK, [ablo]:} & 800 \\ \text{2Sg, SPEAK, [ablas]:} & 200 \\ \text{2Pl, COOK, [kosinais]:} & 100 \\ \text{3Pl, COOK, [kosina]:} & 700 \end{cases} \tag{2.22}$$

More generally, if the frequency of a sublexicon is $|s|$, then its prior probability is given by equation 2.23.

$$p(s) = \frac{|s|}{\sum_s |s|} \tag{2.23}$$

One notable result of the Bayesian character of sublexical morphology is its prediction that in the absence of any known base forms for a lexeme, or if only provided base forms that contain no useful information about which sublexicon might give the lexeme a higher probability, a speaker solving the paradigm cell filling problem will simply sample from the prior distribution over sublexicons in order to infer a derivative form for that lexeme. When base forms that contain information relevant to sublexicon choice are available, they will pull speakers' predicted distributions away from the prior distribution, but in general the prior distribution will always play a significant role in shaping their *a posteriori* distributions, i.e. the distributions the speaker arrives at by combining base likelihoods and prior probabilities. Chapter 4 confirms these predictions and discusses how *regularization* of gatekeeper grammar weights can be used to modulate the relative importance of likelihood terms and prior terms.

2.4.3 Normalization

According to the definition of probability, the individual sublexicon probabilities that make up the distribution $p(S|b_1, b_2, ...b_n)$ must sum to 1.0. This property is guaranteed by dividing each sublexicon's numerator $p(b_1, b_2, ...b_n|s)p(s)$ by a normalization constant Z. As with the constant Z_g from subsection 2.4.1, which normalizes the distribution $p(b_1, b_2, ...b_n|s)$ by summing across the exponentiated harmony values of (an approximation of) all possible base forms, this constant Z constitutes a sum. Specifically, Z is the sum of the products of each sublexicon's $p(b_1, b_2, ...b_n|s)$ and $p(s)$, as expressed in 2.24.

$$Z = \sum_{s} p(b_1, b_2, ...b_n|s)p(s) \tag{2.24}$$

In the case of calculating Z, the terms $p(b_1, b_2, ..b_n|s)$ and $p(s)$ are calculated in the same manner as described in the preceding subsections for the numerator in equation 2.7.

2.4.4 Generating the candidate set

While the attested forms of the bases b_1 through b_n are provided to the model, the set of candidates D must itself be inferred. In sublexical morphology, this process is straightforward. Any arbitrary provided base b of the lexeme in question is provided in turn to each sublexicon, and each paradigm sublexicon uses one of its morphological operations to generate a derivative candidate from that base. The forms generated by this process constitute the derivative candidate set.

To examine how candidates are generated, recall that each sublexicon contains a set of morphological operations, each indexed to a base cell and a derivative cell. In an inference task, the derivative cell is set, and so only the operations whose derivative index accords with that derivative cell are relevant.

At this point, the candidate generation process depends on which base forms are available. Any arbitrary base form can be selected from among this set, since the sublexicon contains a morphological operation from each base cell to the target derivative cell. To generate the derivative candidate, the paradigm sublexicon simply applies this properly base-indexed operation to the selected base form. According to the definition of a sublexicon, regardless of which base cell is chosen,

as long as the appropriate morphological operation is used on that base, the same derivative candidate will obtain.

For example, suppose that a paradigm sublexicon for normative European Spanish contains the operations shown in figure 2.4. Note again that as stated in section 2.3, the exact nature of these operations depends on the formalism used to express them and the learning algorithm; the ones shown here are provided as examples.

Inf→1Pl:	change final [ar] to [amos]
3Sg→1Pl:	change final [a] to [amos]

Figure 2.4: Example morphological operations for an "-ar" sublexicon in normative European Spanish.

Recall the earlier example inference task of determining a distribution over the 1Pl forms of the lexeme TO LOVE when out of its inflected forms only its Inf [amar] and its 3Sg [ama] are known. Either of these forms can be used to generate this paradigm sublexicon's 1Pl derivative candidate. Assuming that the 3Sg [ama] is chosen (arbitrarily), then the 3Sg→1Pl operation is applied to this form, yielding the derivative candidate [amamos]. This result would have been the same if the Inf form [amar] and Inf→1Pl operation were chosen instead.

It is possible in some cases that a morphological operation will be unable to apply to a given base form. Behavior in these situations depends, in the PyParadigms implementation, on user-specified parameter values. By default, for example, the operation *change final [ar] to [amos]* would be able to apply to a base like [asir] by ignoring the operation's mention of the specific sequence [ar] and instead replacing the segments in the same position (word-final, in this example), with the sequence [amos], yielding the derivative [asamos]. Users may specify instead that the operation should require material in the base to match that specified in the operation, in which case this example would not yield a derivative form but instead instantly assign the sublexicon a probability of zero. This outcome would be the same if, for example, an operation altering the second-to-last syllable nucleus of a base were applied to a base with only one syllable. Sublexicons given a probability of zero are effectively treated as non-existent for the remainder of a derivation.

2.4.5 Bringing everything together

The previous subsections have shown, in the abstract and for a running example from Spanish, how sublexical morphology models determine the different values necessary to perform derivative inference. Equation 2.16, repeated here as 2.25, shows how a base likelihood $p(b_1, b_2, ...b_n|s)$, a sublexicon prior probability $p(s)$, the normalization constant Z, and each sublexicon's derivative candidate can be combined to yield a distribution over derivative candidates.

$$\begin{aligned} & p(D|B_1, B_2, ...B_n) \\ &= p(S|B_1, B_2, ...B_n) \\ &= \frac{p(B_1, B_2, ...B_n|S)p(S)}{Z} \end{aligned} \tag{2.25}$$

As a concrete demonstration of how these values relate, consider once again the Spanish example of inferring a distribution over 1Pl forms of TO LOVE given the Inf [amar] and the 3Sg [ama]. The same three sublexicons that have been referred to throughout this section still make up the model: the "-ar" sublexicon, the "-er" sublexicon, and the "-ir" sublexicon. As shown in the subsection about base likelihood, the likelihood of the base forms Inf [amar] and the 3Sg [ama] given each of these sublexicons might be 0.331, 0.003, and 0.001, respectively. Suppose that the respective prior probabilities of these sublexicons are 0.4, 0.4, and 0.2. The term $p(b_1, b_2, ...b_n|s)p(s)$, i.e. the pre-normalization numerator, for each of the three sublexicons would therefore be 0.1324, .0012, and 0.0002, respectively. Z is equal to the sum of these values: 0.1338. Dividing the numerator for each sublexicon by Z gives the *a posteriori* probability of each sublexicon, that is, its probability taking both likelihood and prior probabilities into account: 0.990, 0.009, and 0.001 for the "-ar", "-er", and "-ir" sublexicons.

At this point, we have calculated the distribution on the middle level of equation 2.25. All that remains is to assign these sublexicon probabilities to their appropriate derivative candidates. Suppose that in this case the three sublexicons' morphological operations produce the following three candidates for the 1Pl form: [amamos], [amemos], and [amimos]. None of these derivative forms are homophonous, and so sublexicon probabilities can be assigned to them without any need for sum-

ming probabilities of same-candidate sublexicons. Consequently, the speaker or model would assign a probability distribution of {0.990, 0.009, 0.001} to the three candidates {[amamos], [amemos], [amimos]}. Sampling a single form from this distribution, a speaker solving this particular instance of the paradigm cell filling problem would therefore be very likely to produce [amamos] as the 1Pl form of TO LOVE.

2.5 Learning in sublexical morphology

Sublexical morphology models possess the virtue of demonstrable learnability. Moreover, in contrast with e.g. Network Morphology (Brown & Hippisley, 2012), these models can be learned from phonological forms of words without analyst-provided morph divisions, making its learning inputs more similar to those presumably encountered by human learners. This section describes the learning algorithm for sublexical morphology models implemented in PyParadigms as an example of one practical approach to learning these models. The learning algorithm described here proceeds in three sequential steps, as shown in figure 2.5.

Overview of the learning algorithm:

1. Learning the *mapping sublexicons* to and from each cell
2. Learning the paradigm sublexicons of the inflectional system
3. Learning the gatekeeper grammars for the paradigm sublexicons

Figure 2.5: The three steps of the PyParadigms learning algorithm for sublexical morphology models.

This section describes the learning algorithm step-by-step. The first subsection introduces the inputs assumed by the algorithm, and the following three explain the three steps of learning, following the order set forth in figure 2.5. Runtime analyses are provided as appropriate throughout.

2.5.1 Learning algorithm inputs

The learning algorithm for sublexical morphology models takes as inputs the data shown in figure 2.6.

Learning inputs:

- a training set of form transcriptions and their cell labels
- a set of base-indexed constraints
- settings for various parameters that control, e.g., speed vs. accuracy of learning
- (optionally) a set of phonological feature specifications

Figure 2.6: Inputs to the PyParadigms learning algorithm.

The training data consist of a set of forms along with their cell and lexeme labels (and, optionally, their frequencies). These form-cell-lexeme-frequency bundles are equivalent in form to those shown as sublexicons' associated forms in 2.10 and 2.11; 2.26 and 2.27 below show this form in the abstract case and for the Spanish data from 2.11. This similarity is due to the fact that sublexicons' sets of associated forms are simply these training data divided up among the sublexicons. Frequency is set apart from the cell-lexeme-form tuple here simply to show that it is optional; if unspecified, it defaults to a frequency of 1. These constraints can be supplied in a column-delimited text file.

$$\text{training data}: \begin{cases} \text{cell, lexeme, form:} & \text{frequency} \\ \text{cell, lexeme, form:} & \text{frequency} \\ & \dots \end{cases} \tag{2.26}$$

$$\text{training data}: \begin{cases} \text{1Sg, SPEAK, [ablo]:} & 800 \\ \text{2Sg, SPEAK, [ablas]:} & 200 \\ & \dots \\ \text{2Pl, COOK, [kosinais]:} & 100 \\ \text{3Pl, COOK, [kosina]:} & 700 \end{cases} \tag{2.27}$$

The PyParadigms learning algorithm assumes that the user has provided the phonological constraints used in a model's sublexicons' gatekeeper grammars. The algorithm currently has no ability to induce constraints that might be useful in gate-

keeper grammars, and so having the constraint set be part of the training data allows gatekeeper grammars to be constructed. For now the same set of constraints is used by all gatekeeper grammars (although, of course, with potentially different weights including 0 weights), but this limitation is not an inherent part of sublexical phonology. These provided constraints must have the form of gatekeeper grammar constraints previously specified in 2.14 and exemplified for Spanish in 2.15, except that constraints provided to the learning algorithm do not have pre-specified weights. Below, 2.28 shows this format and 2.29 provides an example from Spanish. Note that as the target sequences are regular expressions, the $ symbol here matches the end of a string, similar to a final # in the more abstract notation used before. Cell names must match those provided in the training data, and target sequences must be interpretable as regular expressions. These constraints can be supplied in a column-delimited text file.

$$\text{input constraints}: \begin{cases} \text{cell:} & \text{target sequence} \\ \text{cell:} & \text{target sequence} \\ \dots & \end{cases} \tag{2.28}$$

$$\text{input constraints}: \begin{cases} \text{3Sg:} & \text{a\$} \\ \text{3Sg:} & \text{e\$} \\ \text{3Pl:} & \text{an\$} \\ \dots & \end{cases} \tag{2.29}$$

The PyParadigms learning algorithm also allows the user to specify numerous other boolean and real-valued parameters that affect learning in one way or another, e.g. regularization coefficients for gatekeeper grammars. For further details about these parameters, see the PyParadigms documentation at https://github.com/bhallen/pyparadigms.

Finally, PyParadigms allows the user to provide a set of phonological feature values. If available, these phonological features enable additional functionality. First, these features can be used to specify morphological operations that mutate feature values, potentially allowing the algorithm to create fewer, more general sublexicons, e.g. by unifying an [e]→[i] operation and an [o]→[u] operation into a single [+syll,-high,-low]→[+high] operation. Second, phonological features allow

the user to specify feature-based constraints, which are then translated into regular expressions for evaluation.

2.5.2 Learning mapping sublexicons

When using the PyParadigms learning algorithm, the process of learning a sublexical morphology model begins by learning all of the *mapping sublexicons* in the training data. As originally mentioned in section 2.3, I have so far mostly been using the term *sublexicon* to refer to the *paradigm sublexicons* which constitute the lexical partitions that form the basis of sublexical morphology. However, a given set of training data not only corresponds to a set of paradigm sublexicons, but also to a *set of sets* of *mapping* sublexicons, and determining these mapping sublexicons is the first step in learning a model. This subsection explains how mapping sublexicons differ from paradigm sublexicons and why learning them is essential to the PyParadigms algorithm.

Recall that a paradigm sublexicon is a set of inflected forms that is morphologically homogeneous: within a paradigm sublexicon, for any pair of forms of the same lexeme in that paradigm sublexicon, there is a single morphological operation that can take one of those forms as input and yield the other form. Notably, the forms in a paradigm sublexicon are able to belong to any cell in the inflectional system, and each paradigm sublexicon has a morphological operation for every ordered pair of cells in the system.

A mapping sublexicon can be thought of as similar to a paradigm sublexicon except in that its base cell and derivative cell are fixed. Therefore a mapping sublexicon is indexed to a single base cell and a single derivative cell. A mapping sublexicon includes only one morphological operation: that from its base cell to its derivative cell. Its associated forms are all in either its base cell or its derivative cell. In sublexical morphology, there is no need for a mapping sublexicon to include a gatekeeper grammar, since mapping sublexicons are only used in order to learn paradigm sublexicons. For instance, 2.30 shows an example mapping sublexicon for Spanish present tense verbs, with 1Sg as the base cell and 3Sg as the derivative cell. One useful way to conceptualize the relationship between mapping sublexicons and paradigm sublexicons is that if a set of paradigm sublexicons share the same operation from some cell x to some other cell y, then all their associated

forms/lexemes will be associated with the same $x \rightarrow y$ mapping sublexicon.

$$\text{mapping sublexicon}: \begin{cases} \text{Base cell:} & 1Sg \\ \text{Derivative cell:} & 3Sg \\ \text{Associated forms:} & \begin{cases} \text{1Sg, SPEAK, [ablo]} \\ \text{1Sg, LOVE, [amo]} \\ \text{3Sg, SPEAK, [abla]} \\ \text{3Sg, LOVE, [ama]} \end{cases} \\ \text{Morphological operation:} & [\text{o\#}] \rightarrow [\text{a\#}] \end{cases} \tag{2.30}$$

Mapping sublexicons in sublexical morphology are based on and formally similar to the *sublexicons* of sublexical phonology (Allen & Becker, in review; Gouskova & Newlin-Łukowicz, 2013), a framework that uses sublexical divisions of pairs of forms (e.g. singular–plural pairs) to encode the phonological subregularities relevant to morphological differences in those form pairs. While the sublexicons in sublexical phonology include gatekeeper grammars and are themselves the end state of learning, the PyParadigms learning algorithm for sublexical morphology uses these mapping sublexicons only as a convenient intermediate step in the process of learning paradigm sublexicons.

When a set of training data includes forms belonging to more than two cells, the data can be parsed into multiple sets of mapping sublexicons. Specifically, a set of mapping sublexicons can be constructed for each ordered pair of cells represented in the training data: a set from every base cell to every derivative cell. For a set of training data comprising forms from all present indicative cells in the normative European Spanish verbal system, for example, a set of mapping sublexicons can be constructed for each of the ordered pairs shown in 2.7.

1Sg	→	2Sg
1Sg	→	3Sg
1Sg	→	1Pl
1Sg	→	2Pl
1Sg	→	3Pl
2Sg	→	1Sg
2Sg	→	3Sg
...		
3Pl	→	2Pl

Figure 2.7: Base–derivative cell pairs among the present indicative cells in Spanish verbs.

In order to learn the paradigm sublexicons of this inflectional system, the Py-Paradigms learning algorithm first learns a set of mapping sublexicons for each of these base–derivative cell pairs. For an inflectional system with n cells, $n^2 - n$ sets of mapping sublexicons must be learned, one for each ordered pair of cells, resulting in a runtime on the order of $O(n^2)$. Each pair's set of mapping sublexicons is learned through a procedure nearly identical to the learning algorithm described by Allen & Becker (in review) except for its omission of the learning of mapping sublexicon gatekeeper grammars. Since the scope of this dissertation covers sublexical morphology rather than sublexical phonology, I will not recapitulate the details of the sublexical phonology learning algorithm here, but will instead focus in the next subsection on how the learned mapping sublexicons are combined to establish a set of paradigm sublexicons.

I do note, however, that the set of sets of mapping sublexicons—and therefore the set of paradigm sublexicons—learned from a collection of forms will depend on the morphological operations learned to map between pairs of sets of forms. While the learning algorithm arrives at these operations by postulating numerous hypotheses about the possible operations and then paring them down to the most parsimonious ones jointly able to account for the available data, numerous parameters of the algorithm determine exactly what restrictions are placed on these oper-

ation hypotheses and thus what sets that will be learned. For a discussion of many of these parameters and their influence on learned sublexicons, I refer readers to Allen & Becker (in review).

2.5.3 Learning paradigm sublexicons

Once the mapping sublexicons for an inflectional system have been determined, it is simple to determine the inflectional system's paradigm sublexicons. Recall that a sublexicon is defined as a subset of inflected forms with uniform morphological behavior. Whereas a set of mapping sublexicons needs only to enforce this uniformity with respect to a single base–derivative cell pair, paradigm sublexicons must enforce it with respect to *all* cell pairs. Consequently, determining paradigm sublexicons will generally divide a language's lexemes into a greater number of sublexicons than would determining the mapping sublexicons for those lexemes for a pair of cells; any base–derivative mapping can introduce a distinction between lexemes that must be incorporated into the lexical partitions of a paradigm sublexicon, but the mapping sublexicon for a particular cell pair may have no need for a division required for some other cell pair's mapping sublexicons.

As an example of this principle, consider again the Spanish present indicative verbal system. We have already established that (ignoring most types of irregularities for expositional purposes) there are three paradigm sublexicons: one for "-ar" verbs, one for "-er" verbs, and one for "-ir" verbs. However, the mapping sublexicons for some base–derivative cell pairs would not need to make this three-way distinction. For both "-er" verbs and "-ir" verbs, 3Pl ends in [-en], contrasting with "-ar" verbs' [-an], and so depending on the rule formalism being used, there would need to be at most two 3Sg→3Pl mapping sublexicons for this cell pair—one for inserting [-en] and one for inserting [-an]. However, the mapping from 1Sg forms, which always end in [-o], to 1Pl forms, which end in [-amos], [-emos], or [-imos], requires a three-way division, and so for the set of paradigm sublexicons to enforce sublexical homogeneity, lexemes in the three classes must be separated into three paradigm sublexicons.

The procedure for learning paradigm sublexicons from a set (one for every ordered cell pair) of sets of mapping sublexicons follows naturally from this principle. Once all sets of mapping sublexicons are learned, the algorithm considers each

lexeme in turn. In the simplest case, in which the training data include exactly one form for each lexeme in each cell, each lexeme's paradigm sublexicon is defined as the combination of its forms' mapping sublexicons for each cell-to-cell mapping, where each mapping sublexicon is represented by a label including its cell pair and morphological operations. After such a record has been made for each lexeme, each lexeme has been assigned to a paradigm sublexicon: the distinctive label for a paradigm sublexicon is the set of the cell-pair-and-operations labels for each of its associated mapping sublexicons. The overall inflectional system's set of paradigm sublexicons is therefore simply the set of all paradigm sublexicons corresponding to at least one of these lexemes. Alternatively, the set of paradigm sublexicons can be thought of as essentially the Cartesian product of mapping sublexicons—specifically, the subset of that product that is attested in the training data. The runtime of this step in the learning algorithm is therefore proportional to the number of lexemes m times the number of cell pairs $n^2 - n$, although the operation required for these mn^2 checks is trivial, amounting only to the addition of a listed value to a set of mapping sublexicon labels.

Note that as of the time of writing, the PyParadigms procedure for learning mapping sublexicons' morphological operations is not guaranteed to produce operations that fit the definition of a paradigm sublexicon: the operations may not "converge" by being guaranteed to produce the same derivative candidate regardless of which base cell is chosen. Addressing such cases is the most pressing issue at hand for this project, although I anticipate the solution will be trivial: such convergence will likely be guaranteed if the learning of mapping sublexicons is skipped entirely in favor of learning paradigm sublexicons directly using an equivalent procedure. It may also be possible to achieve this goal by post-processing of mapping sublexicons. In any case, the remainder of this dissertation assumes paradigm sublexicons that do have convergent morphological operations.

2.5.4 Learning gatekeeper grammar weights

The gatekeeper grammar of each paradigm sublexicon is parameterized by a set of weights, one for each of its constraints. The creation of a gatekeeper grammar therefore amounts to the setting of its constraint weights, specifically to values that maximize the likelihood of the training data in that gatekeeper grammar's paradigm

sublexicon while minimizing the probabilities of data outside that paradigm sublexicon. In this way, the learned constraint weights for a particular paradigm sublexicon allow its grammar to assign a high probability to forms phonologically similar to its associated forms while assigning a low probability to forms that drastically differ from them.

In order to arrive at weights that express phonological generalizations about a paradigm sublexicon, the PyParadigms algorithm treats the task of weight setting as a numerical optimization problem. This characterization generally follows the standard definition of a Maximum Entropy harmonic grammar developed by Goldwater & Johnson (2003), Wilson (2006), and Hayes & Wilson (2008). Each sublexicon's gatekeeper grammar is learned independently from the others, and so the total runtime is on the order of $O(n^2)$ for n cells.

Optimization of a gatekeeper grammar's constraint weights begins with the weights initialized to values sampled independently from a Gaussian distribution with mean 0 and variance 1. Optimization is performed using an iterative process in the gradient descent family of algorithms, which progressively adjusts initial weights so as to continually increase (and ultimately maximize) some *objective function*. The objective function used here serves to maximize the likelihood of the training data subject to some amount of regularization (cf. section 4.1 for further discussion of regularization).

Figure 2.8 shows two visual representations of the training data and state of the gatekeeper grammar for the "-ar" sublexicon in Spanish: one before the gradient descent algorithm has begun optimizing weights, and one in the midst of optimization. While the objective function actually outputs the likelihood of the data given a set of weights, this figure shows a more intuitive proxy. Specifically, these tableaux show the observed frequencies of various forms within a sublexicon on the left side, and the rightmost column shows the counts predicted by the current grammar weights. Minimizing the sum of the absolute values of differences between the various forms' observed and expected counts is tantamount to maximizing the likelihood of the data, and so the figure shows observed and predicted counts, their differences, and the sum of the differences' absolute values. Note also that the zero-valued initial weights are a simplification of the Gaussian-distributed weights actually used by the PyParadigms implementation.

Before optimization:

	observed frequency	*w*=0 3Sg: a#	*w*=0 3Sg: e#	*w*=0 3Pl: an#	*w*=0 3Pl: en#	predicted frequency
3Sg: ama	400	1	0	0	0	160
3Pl: aman	240	0	0	1	0	160
3Sg: kome	0	0	1	0	0	160
3Pl: komen	0	0	0	0	1	160
		Sum of absolute values of differences:				640

Mid-optimization:

	observed frequency	*w*=0.6 3Sg: a#	*w*=-0.6 3Sg: e#	*w*=0.3 3Pl: an#	*w*=-0.3 3Pl: en#	predicted frequency
3Sg: ama	400	1	0	0	0	261.38
3Pl: aman	240	0	0	1	0	193.63
3Sg: kome	0	0	1	0	0	78.72
3Pl: komen	0	0	0	0	1	106.27
		Sum of absolute values of differences:				369.98

Figure 2.8: Tableaux showing the training data for the Spanish "-ar" sublexicon with their observed and predicted frequencies. The tableau on top shows predicted frequencies with all weights set to 0, approximating their state before learning, while the bottom tableau shows the result of some amount of iterative optimization.

The probabilities of the various sublexicons in an inflectional system are combined after the fact into a multinomial distribution over sublexicons (comparable to a *softmax* function; Bishop 2006). Because of this fact, and because I approximate the space of all possible forms using the set of all training and testing forms, the weights of a sublexicon's gatekeeper grammar are actually optimized so as to maximally distinguish that sublexicon's associated forms from the associated forms of other sublexicons. This is why the observed frequencies are as shown in figure

2.8: observed frequencies of forms associated with the sublexicon whose grammar is being learned are set to their actual observed frequencies, while the frequencies of forms associated only with a different sublexicon are set to zero. As the figure exemplifies, then, constraints which evaluate to 1 only on forms outside the current sublexicon will tend toward lower, negative weights, while weights of constraints whose outputs are 1 on forms in the current sublexicon will generally rise accordingly with the frequency of those forms. Notably, however, unlike in a naïve Bayes classifier (Lewis, 1998), constraint weights are sensitive to each other at every step in the iterative learning process, allowing the emergence of complex interactions of constraints.

2.6 Relatedness to other theories

Having concluded this chapter's introduction to the theory of sublexical morphology (and the more general theory of Bayesian morphology), I now turn to the question of how this approach compares to other similar theories. According to the typology of inflectional morphology theories set out by Stump (2001), sublexical morphology is an *inferential-realizational* theory: *inferential* essentially because it makes use of morphological operations rather than morphemes, and *realizational* essentially because the semantic features of a derivative form—i.e. its cell label—are what determine its phonological form rather than vice-versa.

This categorization puts sublexical morphology in the company of several other theories of inflectional morphology. The tradition of inferential-realizational theories of inflectional morphology largely corresponds to the *word-and-paradigm* view of inflectional systems, namely one in which entire words (the *forms* of sublexical morphology) and *cells* comprising an inflectional paradigm are the central units of structure and computation, as opposed to being treated as epiphenomenal. Proposals reflecting inferential-realizational or word-and-paradigm approaches include Matthews (1972), Zwicky (1985), Spencer (1991), Anderson (1992), Aronoff (1994), Beard (1995), Blevins (2006). To contrast the general flavor of these theories with sublexical morphology, I will look in greater depth at one that has been particularly influential: Paradigm Function Morphology (Bonami & Boyé, 2007; Stump, 2001).

Paradigm Function Morphology (Stump, 2001) (see also Bonami & Boyé 2007)

posits that any inflectional system can be defined in terms of a *paradigm function* which takes as its inputs the phonological form of a root and a set of morphosyntactic features (i.e. a cell label) and outputs the inflected form for that root in the specified cell through the application of some number of rules. In this sense, Paradigm Function Morphology, like sublexical morphology, models the paradigm cell filling problem, since a paradigm function is able to generate unfamiliar inflected forms. In principle there is no requirement that this function behave deterministically, and so Paradigm Function Morphology is compatible with the observation that there may be multiple viable forms for a particular root in a particular cell. However, this flexibility is distinct from the explicitly probabilistic nature of sublexical morphology, which both learns from and predicts principled yet noisily gradient morphological behavior.

In other respects, although the differences between sublexical morphology and Paradigm Function Morphology may be clear, their comparative merits are more difficult to evaluate. As mentioned before, the starting point of a derivation in Paradigm Function Morphology is a root form and a target cell (set of morphosyntactic features), whereas in sublexical morphology a derivation requires at least one inflected base form instead of a root. From the standpoint of learnability, sublexical morphology has the advantage: while base forms can simply be observed, roots must be inferred, adding an extra complication to learning. Sublexical morphology also allows a model to tailor its predictions depending on the exact set and shapes of the available base forms, rather than predicting the same outputs for a particular root regardless of the encountered forms themselves. However, one could argue that sublexical morphology is less parsimonious in the sense that it requires inflected forms to be stored in the lexicon for later use in derivations, while Paradigm Function Morphology stores only a root for each lexeme.

Similarly, sublexical morphology and Paradigm Function Morphology differ fundamentally in the nature of their mechanisms for transducing output forms from base/root forms. In sublexical morphology, there is only a single morphological operation for each sublexicon which performs the entire modification from a stored base form to a derivative candidate; paradigm functions allow each derivation to proceed in steps, from one *block* of rules to the next, in order to gradually modify the root until it reaches its output form. In addition to the fact that sublexical

morphology's operations are demonstrably learnable, the actual derivation process of applying an operation in sublexical morphology (as distinct from calculating its probability) is marked by formal simplicity. However, this simplicity comes at a cost. Because there is a rule for every ordered pair of cells, the number of stored rules can quickly multiply for even a moderately sized inflectional system, and moreover, there is currently no way for these morphological operations to capture generalizations like one set of cells' forms building procedurally off of another set's forms, which would be easily captured in the rule blocks of Paradigm Function Morphology. In Japanese, for example, the formation of *-tara* conditionals from past-tense forms uses the same morphological change—insertion of a final [-ra]—regardless of whether the base past tense form is negative or affirmative, is causative or not, etc. Since *past negative* and *past affirmative* are as distinct as any two other cells in sublexical morphology, there is no way to ensure only a single operation is used for all derivations of *-tara* conditionals from past tense forms.

It is the computational learnability and implementability of sublexical morphology that most sets it apart from other theories. To my knowledge, the only other computationally implemented model of the paradigm cell filling problem with its own learning algorithm is Network Morphology (Brown & Hippisley, 2012). This theory is similar to Paradigm Function Morphology, but with abstract *nodes* from which intermediate and output word forms inherit inflectional properties instead of blocks of rules. Network Morphology is implemented using the DATR formal language, and its creators have provided an algorithm for learning DATR representations of inflectional systems. But in addition to its lack of compatibility with noisy or probabilistic inputs or outputs, its learning algorithm requires training forms to be annotated by the analyst with boundaries between roots and affixes.

The use of operations or rules that take inflected forms rather than roots as inputs is rare, but in addition to sublexical morphology, the model of inflectional morphology assumed by the Minimal Generalization Learner (Albright & Hayes, 2002) also takes this approach, as does the theory of sublexical phonology (Allen & Becker, in review; Gouskova & Newlin-Łukowicz, 2013) which inspired sublexical morphology. However, neither of these approaches model entire inflectional systems; instead, both only produce models of individual base–derivative cell pair

relationships.

Research in natural language processing has also touched on the paradigm cell filling problem, most notably in the work of Dreyer & Eisner (2011). While their proposal resembles the approach I take here in that it models productivity in inflectional morphology as sampling from inferred probability distributions over inflected forms, numerous other aspects set it apart from sublexical morphology. Most notably, the Dreyer & Eisner model performs learning in a "mostly-unsupervised" (p. 616) manner, starting from inflected forms without labels for their morpho-syntactic/semantic features, meaning that it is incompatible with sublexical morphology in terms of inputs both to learning and to derivation queries. The graphical structure used to express relationships among cells in inflectional paradigms is also fundamentally different (and more complex) than the one I propose. Additionally, it is not clear that their framework would be able to explicitly model the prior probabilities of surface exponents (see chapter 4) that are crucial to sublexical morphology.

Chapter 3

Inference from multiple bases

The sublexical morphology proposal relies crucially on the equation in 2.7, an instantiation of Bayes's theorem, repeated below as 3.1. Recall that S indicates a variable ranging over sublexicons s, while each B_{cell} indicates a variable ranging over the possible forms or shapes b of a lexeme in a particular $cell$.

$$p(S|B_1, B_2, ...B_n) \propto p(B_1, B_2, ...B_n|S)p(S) \tag{3.1}$$

This chapter and the following one each target a particular, potentially contentious aspect of this equation, supporting its inclusion in 3.1 using novel experimental findings. These results are drawn from *wug* tests (Berko, 1958) on Icelandic and Polish which break new ground by adding a novel element—multiple base forms—to the traditional *wug* test paradigm.

This chapter serves to empirically validate the inclusion of bases B_1 through B_n in the conditional probabilities in 3.1, and also to probe the limits of speakers' abilities to combine information provided by these multiple bases. Intuitively, use of the expression $B_1, B_2, ...B_n$ here denotes that the probability of a particular derivative candidate d (by way of the sublexicon s that generates it; cf. 2.4) depends on the observed shapes of multiple other base forms $B_1, B_2, ...B_n$ of that derivative's lexeme. In other words, according to this hypothesis, speakers can productively use information from multiple known forms of a lexeme when inferring unknown forms of that lexeme. Returning to the normative European Spanish example from the previous chapter, this expression might be used to state that the probability the

grammar gives to some unobserved 1SgPresIndic verb candidate (say, [pwento]) as opposed to some other candidate for that form (say, [ponto]) can differ depending on whether the speaker knows the 2SgPresIndic form of that lexeme, or the 3SgPresIndic form of that lexeme, or both.

This position contrasts with a view of morphological inference in which an inflectional system has a single *privileged* base cell such that speakers' inferences about other inflected forms can only make use of information in that privileged cell's base form(s). If only information in the privileged base is used in predicting a derivative form, then the derivative is conditionally independent of the non-privileged bases given the privileged base. Using the definition of conditional independence, this restriction can be written as shown in equation 3.2, where $B_1, B_2, ...B_n$ includes $B_{privileged}$. I use D rather than S here to signify that this restriction in no way depends on the assumption of sublexical morphology.

$$p(D|B_1, B_2, ...B_n) = p(D|B_{privileged}) \tag{3.2}$$

In other words, this view assumes that the probability of a derived form of a lexeme given any of its other forms is equal to the probability of that derived form given only the form of its privileged base. Less formally, this equation states that no information contained in forms of a lexeme other than the privileged base form can exert any influence on probabilities given to that lexeme's derived forms.

While this more restricted hypothesis about inflectional systems may appear unnecessarily conservative, it is based on a related hypothesis that has proved useful for understanding and modeling mechanisms of historical morphological change: Albright's (2002 *et seq.*) *single surface base hypothesis*. Because of the empirical successes of the single surface base hypothesis—and because the sublexical view of morphology derives from a similar set of assumptions about inflectional morphology—I consider it worthwhile to test whether the single surface base hypothesis is tenable in a probabilistic model of inflectional morphology.

This chapter proposes a probabilistic interpretation of the single surface base hypothesis in section 3.1, adapting it to the language and formalisms used in this dissertation. Using data from Icelandic (section 3.2), I then conclude on the basis of experimental evidence presented in section 3.3 that contrary to the single sur-

face base hypothesis, multiple base forms must be available and usable in inflectional inference even in the inference of a single derivative form. This conclusion supports the inclusion of the expression $B_1, B_2, ...B_n$ in equation 3.1. Following up on these results, section 3.4 addresses the question of whether the expression $p(B_1, B_2, ...B_n|S)$ can be decomposed to facilitate learning and inference, and it presents an experimental study of Polish noun declensions designed to answer this question. These results tentatively suggest that speakers may be systematically limited in how they can combine information from multiple bases, resulting in a picture of speaker capabilities that simplifies the modeling task for linguists. Overall, then, the chapter's empirical results support the inclusion of multiple bases in equation 3.1 and shed light on the modes of interaction among these bases.

3.1 Single-base hypotheses

Albright (2002, p. 11) defines the *single surface base hypothesis* as a proposal that:

> ...for one form in the paradigm (the [privileged] base), there are no rules that can be used to synthesize it, and memorization is the only option [for speakers to be able to produce this form]. Other forms in the paradigm may be memorized or may be synthesized, but synthesis must be done via operations on the [privileged] base form...

For the purposes of this dissertation, the single surface base hypothesis can be summarized as comprising two claims: (a) that any unknown inflected form of a lexeme must be generated only from the memorized form of that lexeme in a single cell, and (b) that the particular *privileged* cell used for this generation process is the same across all lexemes in the language and all possible derivative cells. The cell privileged in this way is the cell whose forms have the fewest neutralizations of morpho-phonological contrasts among inflected forms, i.e. the cell with the highest predictive value as a base from which other forms can be inferred.

As an example of this hypothesis at work, consider one locus of historical change in inflection in High German as described by Albright (2008). Figure 3.1 shows three classes of nouns in Middle High German. The top class exhibits a [x] $\sim$ [∅] alternation attributable to intervocalic loss of the ancestor of [x]. This change resulted in a neutralization in the NomPl of the top class and non-alternating

middle class. The bottom class has [x] at the end of NomSg forms and at the end of the stem in NomPl forms, which is due to a sound change of [k] to [x] after the elimination of intervocalic [x]. This second change resulted in a neutralization of the bottom and top classes in the NomSg. The patterns of vowel umlaut are not relevant to this example.

NomSg	NomPl	Gloss
flox	flœːe	‘flea’
rex	reː(j)e	‘deer’
kuː	kyːe	‘cow’
weː	weː(j)e	‘woe’
kox	kœːxe	‘cook’
pex	pexe	‘pitch’

Figure 3.1: Examples of three classes of nouns in Middle High German, in the NomSg and NomPl. Examples are reconstructions of historical forms, transcribed using the IPA.

In modern High German, however, the top and middle classes have been neutralized in the NomSg, with all such forms being vowel-final, e.g. [floː] ‘flea’. According to the single surface base hypothesis, this neutralization—a form of paradigm leveling—results from the fact that speakers are only able to make inferences from the NomPl form, which best maintains class contrasts overall (aside from the neutralizations shown here) and therefore is the privileged base. Therefore any inference about the shape of a lexeme’s other forms must rely only on information present in its NomPl form, and so contrasts which are neutralized in that form, such as that between the top and middle classes here, are at risk of being lost across the entire paradigm. This restriction illustrates part (a) of the single surface base hypothesis. Moreover, according to part (b) of the hypothesis, the privileged cell must be the cell that neutralizes the fewest contrasts in the inflectional system overall (the NomPl in High German), even if it is not especially predictive of forms in the particular derivative cell whose form is being inferred, as in this example. Indeed, these predictions are borne out in the modern NomSg forms of

words like [floː] ‘flea’, where contrasts lost in the NomPl are extended to NomSg forms, showing the explanatory power of the two parts of the single surface base hypothesis together.

Note that part (b) of the single surface base hypothesis further restricts part (a). In other words, while it is possible to propose a weaker version of the hypothesis including part (a) but excluding part (b)—i.e. a view in which only one base form can be used for generating derivatives but in which the choice of base form can vary—part (b) logically requires part (a).[1] In this section, I will review the motivations for both parts of the hypothesis. I will then consider how the hypothesis might be adapted to a probabilistic setting.

3.1.1 Motivations for the single surface base hypothesis

Assuming the validity of the single surface base hypothesis benefits a word-based theory of inflectional morphology in two ways, both of which are relevant to the aims of this dissertation. First, restricting the ways that novel forms can be generated may reduce the computational and representational complexity of the morphological grammar. Second, the hypothesis accurately predicts patterns of paradigm leveling in multiple languages. The remainder of this section focuses on the first of these—the way that having a single privileged base can reduce model complexity—using this opportunity to introduce a graph-theoretic interpretation of inference in word-based inflectional morphology. The phenomenon of paradigm leveling, which I used in the previous subsection to explain the single surface base hypothesis, is covered in more detail in section 5.2, where I demonstrate that sublexical morphology is able to account for such patterns even without the single surface base hypothesis.

In a word-based model of inflectional morphology, specifically one like that assumed by Albright (2002) in which novel wordforms must be generated through the application of morphological operations to other wordforms, the grammar must contain information about how the speaker can generate a wordform in any one cell

[1] One could also propose versions of the single surface base hypothesis with an arbitrary number of privileged cells, e.g. one in which there are exactly two privileged cells whose forms must be memorized and which are the only two cells whose forms can be used for inference. This generalization shades into the concept of *principal parts* (Stump & Finkel, 2013). Due largely to considerations related to my experimental methodology, I focus here on the most restrictive hypothesis, one with only a single privileged base.

from the same lexeme's wordform in any other cell. A primary purpose of the single surface base hypothesis is to simplify the information about an inflectional system that must be stored in the speaker's grammar by restricting the "paths" through which a novel form can be derived, reducing the grammar's formal complexity and presumably facilitating learnability. (Note, however, that the single surface base hypothesis itself provides no explanation for how speakers would learn the identity of the privileged base cell.)

To observe these two effects of the hypothesis, we can contrast two views of how an unknown inflected form is derived from other known forms of the same lexeme, one adhering to the single surface base hypothesis and the other taking the opposite extreme. For both of these, I will make use of a graph-theoretic representation of inflectional systems in which labeled nodes represent (wordforms in) cells in the system and edges (directed, i.e. one-way, or undirected, i.e. bi-directional) represent the "paths" via which a wordform in one cell can be derived from that in another cell. Further extending the spirit of the single surface base hypothesis into this graph-theoretic idiom, I assume that the process of a derivation can traverse only a single edge—i.e. that there are no intermediate representations.

One such graph structure for an inflectional system with four cells is a *complete* graph with undirected edges, as shown in 3.2.

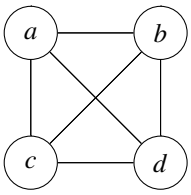

Figure 3.2: A fully connected inflection graph. Each form in the inflectional system can be used to generate each other form.

Such a view of the paths of derivation in inflectional morphology is maximally unconstrained. For any given target cell (*a*, *b*, *c*, or *d* in figure 3.2), the form in any other cell can be used to generate the form in the target cell. Each edge in the graph represents a function that generates a form in one cell—or a set of candidates for that form—from the form in another cell.

However, this freedom comes at the cost of substantial computational complexity. For an inflectional system with n nodes, the number of edges in its complete graph is $n(n-1)/2$. Moreover, each edge represents two mappings: one from node/cell x to node/cell y, and one from y to x. Thus the number of morphological operations deriving one cell's form from another cell's form is double this amount, $n(n-1)$. In the absence of disconfirming evidence, it would be computationally and theoretically preferable to use a less information-dense representation for the grammar.

According to Albright's (2002, 2008 *et seq.*) single surface base hypothesis, the number of morphological operations (edges in a graphical representation) deriving one inflected form from another amounts to only $n-1$, where n is again the number of cells in the inflectional system. Figure 3.3 gives a graphical representation of the possible derivations in a four-cell inflectional system under the single surface base hypothesis; note that cell a is assumed to be the privileged base, and that edges are directed because the a form must be memorized and cannot be generated from other forms.

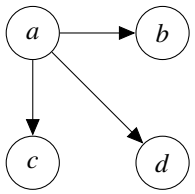

Figure 3.3: An inflection graph under the single surface base hypothesis, with cell a as the privileged base. The form in a must be memorized, and other forms must be derived only from a.

Albright (2002, p. 9) also discusses an intermediate alternative according to which "each [cell] in the paradigm must be derived from at most one unique base, but different [cells] may be derived using different bases." Figure 3.4 shows an example of this kind of system. While this proposal lacks the strictness of the single surface base hypothesis, it still predicts that any particular cell's form can be generated only from a single other cell's form. This version of the hypothesis amounts to an assumption of its part (a) and a rejection of its part (b).

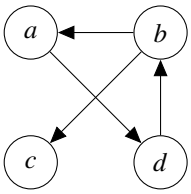

Figure 3.4: An inflection graph under a weakened version of the single surface base hypothesis, assuming that each cell can be generated from some cell.

The constant thread across all versions of the single surface base hypothesis proposed by Albright remains the restriction that wordforms in any particular cell can only be generated from a single other cell (or else memorized). Despite the advantages in formal complexity that these hypotheses afford, I conclude in this chapter that ultimately—and unfortunately, from the standpoint of computational and formal complexity—the only empirically valid model of inference in inflectional morphology violates these hypotheses, necessitating a grammar that looks more like the complete graph in 3.2 than any of the simpler graphs in this section.

3.1.2 A probabilistic single surface base hypothesis

The single surface base hypothesis itself deals only with the question of which base cell's forms can be used to synthesize, i.e. generate, candidates for inflected forms. In literature on the single surface base hypothesis (Albright, 2002, 2008; Albright & Hayes, 2003), determining which of these candidates will be uttered relies on the machinery of the Minimal Generalization Learner, a framework for learning and applying sets of morphological rules for inflectional morphology. As Albright & Hayes (2002) in particular makes clear, while the *confidence* measures used by the Minimal Generalization Learner allow quantitative comparison of rules and therefore of candidates, the Minimal Generalization Learner does not learn probabilistic grammars. In order to test the spirit of the single surface base hypothesis in the specific context of a probabilistic view of the paradigm cell filling problem, it is necessary to formulate an explicitly probabilistic version of the hypothesis.

In the language of probability theory, we can consider the form of a derived inflected word to be a discrete random variable D, such that $p(D)$ indicates a distri-

bution of probability mass across a finite set of candidate wordforms. For example, 3.3 shows a possible set of candidate wordforms for the plural form of the English lexeme GIRAFFE and a possible probability distribution over them. According to the definition of probability, the probabilities across all candidates must sum to 1.

$$\text{p(D)}: \begin{cases} \text{p(D = [dʒɪræfs])} & = 0.7 \quad \text{(cf. LAUGH)} \\ \text{p(D = [dʒɪrævz])} & = 0.297 \quad \text{(cf. CALF)} \\ \text{p(D = [dʒɪræfɪz])} & = 0.002 \quad \text{(cf. CLASS)} \\ \text{p(D = [dʒɪrævɪz])} & = 0.001 \quad \text{(cf. HOUSE)} \end{cases} \tag{3.3}$$

I propose that the appropriate probabilistic interpretation of the single surface base hypothesis is one that prohibits non-privileged base forms from affecting the calculation of probability distributions over derivative candidates. As described in the introduction to this section, then, we can use the definition of conditional independence to arrive at the probability-theoretic version of the single surface base hypothesis given by the equation in 3.2, repeated below as 3.4.

$$p(D|B_1, B_2, ...B_n) = p(D|B_{privileged}) \tag{3.4}$$

In other words, the probabilistic version of the hypothesis predicts that the probability of a derived form of a lexeme given all of its other forms equals the probability of that derived form given only the form of its privileged base. Less formally, this equation states that no information contained in forms of a lexeme other than the base form can exert any influence on probabilities given to that lexeme's derived forms.

Note that this probabilistic formulation of the single surface base hypothesis encompasses only its part (a) as defined at the start of this section: the limitation of making use of only one base form. The definition in equation 3.4 is agnostic with respect to the question of whether that privileged base cell must be the same across all lexemes and all derivative cells in an inflectional system, i.e. part (b) of the definition from the beginning of this section. The requirement for an inflectional system to have only one cell ever used as the privileged base needs no probabilistic reinterpretation and can still be applied unchanged in cases where the

grammar predicts a probability distribution over derivative candidates. Moreover, it is still the case in a probabilistic setting that this restriction (b) logically requires its part (a), i.e. equation 3.4, and so by falsifying this equation, the limitation to one privileged base across an entire inflectional system can also be falsified.

This discussion brings us to the question of how to falsify the probabilistic version of the single surface base hypothesis. As usual in hypothesis testing, it is necessary to determine what falsifiable predictions the hypothesis makes. To this end, we can rewrite equation 3.4 to specifically indicate that the privileged base is one of the bases in $B_1, B_2, ...B_n$, making it clear that the conditioning factor on the right hand side of the equation is a proper subset of those on the left:

$$p(D|B_{privileged}, B_{other}, ...B_n) = p(D|B_{privileged}) \tag{3.5}$$

In other words, it is possible to test the probabilistic single surface base hypothesis by determining whether there exist inflectional systems such that bases other than the privileged base can affect the probability of a derivative candidate. This type of conjecture—proving that some equality does not hold, especially in cases where one condition is a superset of another—lends itself to scrutiny by experiment. The following section describes an experiment designed to test the very equality in 3.5.

3.2 Icelandic nouns and multiple bases

In previous sections, this chapter has made a case for the usefulness of the single surface base hypothesis, and it has also proposed a way that the hypothesis might be adapted to a probabilistic model of inflectional morphology. In these two final sections, however, I present evidence that the single surface base hypothesis as described here cannot hold true for all languages. This discussion focuses on a particular pattern in Icelandic noun declensions that appears to require information from multiple bases in the production of a single derivative form. To expand this argument from the domain of abstract properties of an inflectional system to concrete facts of speaker behavior, I report experimental evidence that Icelandic speakers do indeed make use of multiple base forms in such situations, counter to the predictions of the single surface base hypothesis.

3.2.1 Icelandic noun inflection

Nouns in Icelandic inflect for case and for number (Einarsson, 1949; Kress, 1982). Having four cases and two numbers, the inflectional system of nominals comprises a total of eight cells. These eight case–number combinations are shown in Figure 3.5 along with their abbreviations.

	Nominative	Accusative	Dative	Genitive
Singular	*NomSg*	*AccSg*	*DatSg*	*GenSg*
Plural	*NomPl*	*AccPl*	*DatPl*	*GenPl*

Figure 3.5: The four cases and two numbers of Icelandic nouns, as well as their abbreviations.

The phonological exponents of case and number information primarily consist of suffixes which jointly express case and number, for example, a DatPl suffix written as *-um* and pronounced [ʏm], or an orthographically and phonologically null NomSg suffix. Icelandic nouns sometimes also exhibit stem vowel alternations, e.g. NEED NomSg *þörf* [θœrv] ∼ NomPl *þarf-ir* [θarvir]. While such vowel alternations present no issues for the models I advocate, for the sake of simplicity, I focus here on suffix patterns. Note in addition that nouns can be marked for definiteness, e.g. HOUSE DatPl definite *hús-unum* [huːsɔnʏm] ∼ DatPl indefinite *hús-um* [huːsʏm], but that the domain of immediate interest is limited to indefinite forms. For the remainder of this discussion of Icelandic, I will use orthographic representations rather than transcriptions, as the experimental data I will present relates only directly to the former. Orthographic representations can be fairly viewed as roughly corresponding to equivalent IPA symbols, and more fine-grained details of the Icelandic writing system and its relation to segmental phonology are not relevant to this discussion.

Most case-number combinations exhibit a variety of suffixal exponents. These suffixes combine to form what have traditionally been called *inflectional classes*, i.e. sets of lexemes with identical inflection across the entire paradigm (Müller, 2005). These classes also interact with grammatical gender: inflectional class

and gender (masculine, feminine, or neuter) are strongly correlated. For example, Figure 3.6 shows the suffix co-occurrence patterns of six inflectional classes whose member lexemes are exclusively (except for one or two proper names of the STANZA class; Gunnar Ó. Hansson, p.c.) of the feminine gender.

	CHIP	NYMPH	BAY	HEATH	MOVEMENT	STANZA
NomSg	*flís*	*dís*	*vík*	*heið-i*	*hreyfing*	*vís-a*
AccSg	*flís*	*dís*	*vík*	*heið-i*	*hreyfing-u*	*vís-u*
DatSg	*flís*	*dís*	*vík*	*heið-i*	*hreyfing-u*	*vís-u*
GenSg	*flís-ar*	*dís-ar*	*vík-ur*	*heið-ar*	*hreyfing-ar*	*vís-u*
NomPl	*flís-ar*	*dís-ir*	*vík-ur*	*heið-ar*	*hreyfing-ar*	*vís-ur*
AccPl	*flís-ar*	*dís-ir*	*vík-ur*	*heið-ar*	*hreyfing-ar*	*vís-ur*
DatPl	*flís-um*	*dís-um*	*vík-um*	*heið-um*	*hreyfing-um*	*vís-um*
GenPl	*flís-a*	*dís-a*	*vík-a*	*heið-a*	*hreyfing-a*	*vís-na*

Figure 3.6: Representative words and their suffix paradigms from six inflectional classes associated with the feminine gender.

Information about an inflected form's suffix provides information about what inflectional class that form's lexeme could belong to—and, more descriptively, provides information about what suffixes other inflected forms of that lexeme are likely to take. Knowledge of a lexeme's gender can also provide information about its likely suffixes (and vice-versa). For example, according to the classes shown in Figure 3.6, if it is known that some arbitrary lexeme is feminine and that its NomPl takes the *-ir* suffix, then one can infer that its GenSg takes the *-ar* suffix (as a lexeme in the NYMPH inflectional class). Conversely, if it is known that a lexeme's GenSg takes the *-s* suffix (common for masculines and neuters, but not used for feminines), then one can infer that it is not of the feminine gender. The primary goal of the experiment described in the next section is to determine whether native speakers of Icelandic in fact use knowledge of a lexeme's suffixes to perform inference in this way.

3.2.2 Predictors of the Icelandic AccPl

For the purposes of setting up the experiment introduced in the next section, I focus now on the AccPl forms of Icelandic nouns. In particular, the discussion will focus on the four AccPl suffixes described in Figure 3.7, as these suffixes' distributions comprise a basis for testing the single surface base hypothesis.

Suffix	Gender	Theme Vowel
-a	masculine	a
-i	masculine	i
-ar	feminine	a
-ir	feminine	i

Figure 3.7: Four AccPl suffixes of Icelandic nouns, their usual genders, and their stem vowels.

Of central interest is the question of how a noun's AccPl suffix can be predicted given knowledge only of its GenSg and/or NomPl forms, that is, without any additional information about gender or other inflected forms. The GenSg and NomPl are classically considered the "principal parts" of Icelandic nouns, owing to their great predictiveness of other cells' inflected forms, and are standardly listed in dictionary entries, e.g. Árnason (2007). The GenSg of a lexeme provides information about its gender and therefore also its AccPl: a GenSg with the *-s* suffix is compatible with the masculine AccPl suffixes *-a* and *-i* but not with *-ar* or *-ir*, while the *-ar* GenSg suffix is compatible primarily with the feminine AccPl suffixes *-ar* and *-ir*. A lexeme's NomPl form provides complementary information: a NomPl with the *-ar* suffix is compatible only with the a-stem AccPl suffixes *-a* and *-ar*, while the *-ir* NomPl suffix is compatible only with the i-stem AccPl suffixes *-i* and *-ir*.

Figure 3.8 illustrates these patterns by showing counts (i.e. type frequencies) of Icelandic nouns with any one of the four GenSg-NomPl-AccPl suffix constellations described above, drawn from the *Database of Modern Icelandic Inflection* (Bjarnadóttir, 2012) and grouped according to their AccPl, GenSg, and NomPl suffixes. These counts constitute all three forms from a total of approximately 182,000 lex-

emes. Numbers in parentheses provide more abstract estimates of how much of the lexicon falls into each category by conflating all compounds containing the same head noun, e.g. counting *vorlaukur* 'spring onion' (lit. 'spring-onion') and *graslaukur* 'chives' (lit. 'grass-onion') together as having a frequency of one.

	AccPl *-a*	AccPl *-i*	AccPl *-ar*	AccPl *-ir*
GenSg *-s*	**17,193**	0	0	0
NomPl *-ar*	**(1,880)**	(0)	(0)	(0)
GenSg *-s*	0	**3,289**	0	0
NomPl *-ir*	(0)	**(153)**	(0)	(0)
GenSg *-ar*	318	0	**6,704**	0
NomPl *-ar*	(14)	(0)	**(476)**	(0)
GenSg *-ar*	0	3,983	0	**14,274**
NomPl *-ir*	(0)	(108)	(0)	**(714)**

Figure 3.8: Raw counts (and head noun-based counts) of Icelandic noun forms grouped by their AccPl, GenSg, and NomPl suffixes, taken from the *Database of Modern Icelandic Inflection* (Bjarnadóttir, 2012).

As Figure 3.8 shows, knowing both a lexeme's GenSg suffix and its NomPl suffix should allow a speaker to narrow down the four AccPl suffix options to just a single choice—those shown in bold—compatible with implicational relationships in the lexicon, either absolutely (when GenSg is *-s*) or with high certainty (when GenSg is *-ar*).

When only the GenSg suffix or only the NomPl suffix (but not both) are known, according to the same lexical patterns as above, one can perform useful inference about the AccPl suffix, narrowing its viable suffix options from four to fewer. However, knowing either the GenSg or NomPl alone does not suffice to pick out a single highly likely AccPl suffix. Figure 3.9 demonstrates the relevant lexical patterns in such cases of limited useful information.

	AccPl *-a*	AccPl *-i*	AccPl *-ar*	AccPl *-ir*
GenSg *-s*	17,193 (1,880)	3,289 (153)	0	0
GenSg *-ar*	318 (14)	3,983 (108)	6,704 (476)	14,274 (714)

	AccPl *-a*	AccPl *-i*	AccPl *-ar*	AccPl *-ir*
NomPl *-ar*	17,511 (1,894)	0	6,704 (476)	0
NomPl *-ir*	0	7,272 (261)	0	14,274 (714)

Figure 3.9: Raw counts (and head noun-based counts) of Icelandic noun forms grouped by their AccPl, GenSg, and NomPl suffixes, but with GenSg-based and NomPl-based groupings performed separately, taken from the *Database of Modern Icelandic Inflection* (Bjarnadóttir, 2012).

The schematic in Figure 3.10 provides an abstracted view of the lexical facts set forth so far. It displays the four AccPl suffixes of interest as lying in a table with four sections. With high accuracy, knowledge of a lexeme's GenSg predicts which *column* in the table should contain that lexeme's AccPl suffix, while knowledge of its NomPl predicts which *row* should contain its AccPl suffix.

		GenSg	
		-s	*-ar*
NomPl	*-ar*	*-a*	*-ar*
	-ir	*-i*	*-ir*

Figure 3.10: Schematic of four possible AccPl suffixes in Icelandic, with their typical lexical correspondences to GenSg and NomPl suffixes.

This situation, in which two base forms provide complementary information conceivably usable in inferring a derivative form, provides an ideal case for testing the probabilistic formulation of the single surface base hypothesis. The next section describes an experiment designed to test whether information in both base forms is accessible to Icelandic speakers for use in inference (part b of the hypothesis) and whether speakers can combine information from the two forms in order to improve

the accuracy of their inference (part a of the hypothesis).

3.3 Falsifying the single-base restriction: an Icelandic experiment

As the previous section describes, certain inflectional morphology tasks plausibly faced by native speakers of Icelandic—such as the prediction of a noun's AccPl form in the absence of gender information—exhibit patterns of incomplete predictability which would require that an analyst considering the pattern look beyond the information in just a single base form. The purpose of this section is to determine whether native speakers of Icelandic adjust their judgments in such situations by considering, as analysts can, information from multiple base forms. The finding that native speakers can do so would directly contravene the single surface base hypothesis, and indeed this result obtains in the experiment presented here.

3.3.1 Methodology

A total of 191 Icelandic native speaker participants were recruited by posts on mailing lists and social media and through word of mouth. Of these participants, 123 completed the entire experiment and described themselves as native speakers of Icelandic, and it is these 123 participants' data which I analyze here. Among these participants, approximately 30% described themselves as having been born between 1948 and 1957, and approximately 31% as having been born between 1958 and 1967; the 1938–1947, 1948–1957, and 1978–1987 ranges each comprised approximately 10% as many participants, with the remaining few participants either born before 1937, born after 1988, or declining to specify. Among these same 123, approximately 68% described their gender as female, approximately 26% as male, and approximately 1.6% as another gender, while the rest declined to specify.

Figure 3.11: A screenshot of one trial frame in the Icelandic experiment. This frame has presented the DatPl and NomPl forms of the GLEIT nonce lexeme and is now eliciting a choice for its AccPl.

Procedure

The experiment itself was carried out entirely online using the Experigen (Becker & Levine, 2012) framework. On the Experigen web interface, participants were given information about the experiment and were then provided a consent form to electronically sign. Each participant who consented engaged next in two non-randomized practice trials, after which she or he completed thirty-two test trials. Finally, all consenting participants filled out a demographic questionnaire asking for non-identifying personal information. All parts of the experiment were presented in Icelandic, as translated from English by a native speaker of Icelandic. The English translation of this questionnaire can be found in the appendix.

Each trial concerned a single novel lexeme designed to resemble existing Icelandic noun lexemes. A participant would first be exposed to some number of inflected forms of the novel lexeme in carrier sentences. The task was then to

select one preferred AccPl form of the lexeme out of a fixed set of four options, one created using each of the four AccPl suffixes discussed in the previous section. The order of presentation of these suffix options was randomized. The participant's choice of AccPl form was recorded as the key experimental measure.

Primary manipulation

The information about a novel lexeme presented to a participant before the AccPl choice task varied according to each stimulus frame's *presentation condition*. This variable ranged across the four options shown below, which correspond to the inflected forms shown to participants before they were asked to select an AccPl form. The DatPl form conveys no information about which AccPl suffix is most appropriate, because it takes an *-um* suffix for all nouns in the language (with the exception of a handful of monosyllabic vowel-final stems). The DatPl form is always provided to introduce participants to each novel lexeme and encourage them to think of it as an existing Icelandic noun. Some trials then also present either the GenSg or NomPl, which—according to the lexical patterns described in the previous section—provides limited useful information about what the AccPl form should be. Some other trials then provide the remaining base form, which—again, according to the lexical patterns described in the previous section—should leave only one AccPl exponent likely.

1. DatPl only [uninformative]
2. DatPl, then GenSg [somewhat informative]
3. DatPl, then NomPl [somewhat informative]
4. DatPl, then GenSg, then NomPl [maximally informative]

Figure 3.12: The four presentation conditions of the Icelandic experiment.

This manipulation tests both parts of the single surface base hypothesis. By using these four presentation conditions, it was possible to evaluate whether both the GenSg form and the NomPl form were used by speakers in inference by comparing responses in conditions 2 and 3 to responses in condition 1. This design also

made it possible to evaluate whether the combined information from the GenSg and NomPl together in condition 4 was used to further modify judgments compared to conditions 2 and 3.

Stimuli

Novel lexeme stems, 32 in total, were assigned randomly to the four inflectional classes corresponding to the four AccPl suffixes forming the range of participants' choices. For clarity, these inflectional classes are repeated in Figure 3.13. These stems were designed so as to minimally influence judgments about their appropriate inflectional class or gender, or at least to balance stems more likely to be judged as belonging to one class/gender with stems more likely to be judged otherwise. Specifically, all stems were of the shape ((C)C)CVC, with their vowels and final consonants drawn from the sets {e, ei, a, ó} and {p, t, m, n}, respectively. These choices of stem shape, vowels, and consonants are based largely on research by Hansson (2006), who shows experimentally that Icelandic speakers are sensitive to lexical correlations among stem shape, gender, and inflection reported by Jónsdóttir (1989, 1993). Two stems were generated from each combination of vowel and final consonant, yielding a total of 32 stems. The two stems in each pair with the same vowel-consonant combination were always assigned to different inflectional classes. To minimize the risk of stems being similar enough to existing stems that the existing lexemes' inflection could unduly affect judgments about the novel stems, I implemented a script that rejected any stems with an edit distance of 1 from any existing stem in the Icelandic lexicon (Bjarnadóttir, 2012), and a linguistically savvy native speaker of Icelandic also performed a similar form of filtering using his own judgments of similarity. A complete list of stimuli can be found in the appendix.

	GenSg	NomPl	AccPl
1 (Masc-a)	*-s*	*-ar*	*-a*
2 (Masc-i)	*-s*	*-ir*	*-i*
3 (Fem-a)	*-ar*	*-ar*	*-ar*
4 (Fem-i)	*-ar*	*-ir*	*-ir*

Figure 3.13: The suffixes of the four inflectional classes into which novel lexeme stems were randomly distributed. All lexemes took the DatPl suffix *-um*.

(Pseudo-)randomization of stimuli was performed in two ways. The order of novel lexemes was first itself randomized. The assignment of lexemes to presentation conditions also varied across participants, but rather than being randomized, Experigen cycled through four possible sets of pairings of lexeme and presentation condition from each participant to the next. Each of these sets of pairings distributed lexemes of various inflectional classes evenly across the four presentation conditions, such that there were always two lexemes in each inflectional class in each presentation condition. The two stems for each vowel-consonant combination were always presented in separate presentation conditions.

Carrier sentences

The DatPl, GenSg, and NomPl forms, when presented, were always given inside a carrier sentence designed to make it clear which paradigm cell the presented form belongs to. These sentences were carefully designed so as not to provide any additional information about gender. The AccPl was also elicited using a carrier sentence, which provided a long underscore (blank line) in the position of the requested AccPl form and which also provided no extra information about gender. After the presentation of each inflected form, participants pressed a button to trigger presentation of the next inflected form or finally the AccPl choice task, always leaving all previously presented inflected forms of the trial lexeme visible in their frame sentences. This button-based procedure was added to encourage participants to consider the information provided by each inflected form. Inflected forms themselves were shown in boldface to make them stand out against the carrier sentences.

Together, all of the carrier sentences formed a short narrative about a person named Jón and his fondness of certain collectible objects, intended to allow participants to think of novel lexemes as the names of obscure trinkets. The narrative was coherent regardless of which sentences were included. The purpose of this design choice, in addition with the explicit instruction before the test trials to think of novel words as real but rare Icelandic nouns, was to encourage participants to use their knowledge of inflectional patterns in existing Icelandic words when performing the AccPl choice task. All presentation and elicitation was performed using orthographic representations, and no recordings were played or made at any time. Frame sentences are provided in Icelandic with English translations in the appendix.

3.3.2 Results

The Icelandic-speaking participants demonstrated, as a group, their ability to make fruitful use of information contained in both GenSg and NomPl forms, as well as their ability to combine information contained in these two base forms. Crucially to these conclusions, participants achieved a higher rate of success in selecting a lexeme's associated AccPl form when provided either its GenSg form or its NomPl form, as compared to when provided only its DatPl form. Moreover, success rates were higher still when participants were provided both the GenSg and the NomPl form of a lexeme. Figure 3.14 summarizes these response patterns.

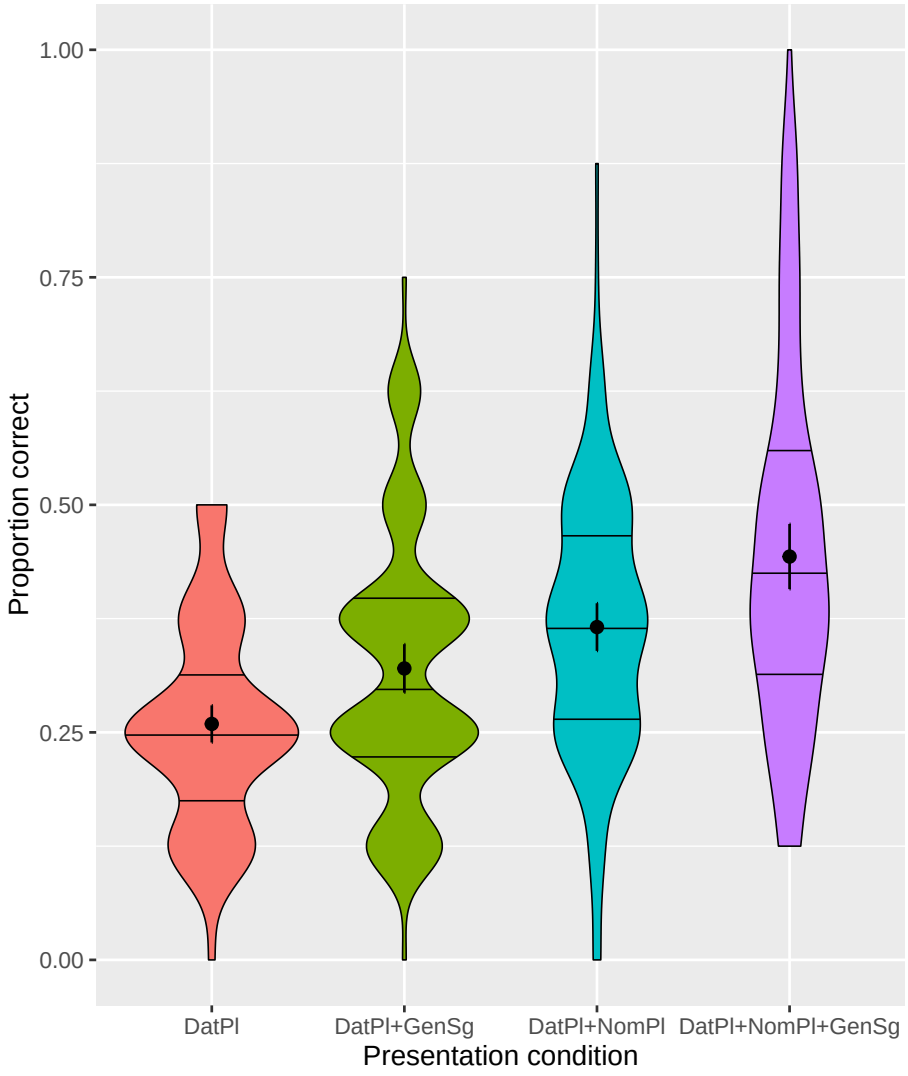

Figure 3.14: Participants' proportions of "correct" responses in the Icelandic experiment by presentation condition. Vertical bars show 95% confidence intervals, and horizontal bars show quartile values. Color-coded probability density functions show kernel estimations of the underlying distributions.

In order to demonstrate the statistical validity of these conclusions, I fit linear models to the experimental data and evaluate the model parameters optimized to describe the data. To this end, I first introduce the linear model framework itself and then demonstrate that according to this model of the responses, neither part of the probabilistic single surface base hypothesis is consistent with the experimental data.

Linear models: set-up

Results from this experiment were analyzed in R (R Core Team, 2013) using the `lme4` package (Bates *et al.*, 2015b), which implements generalized linear mixed effects models (GLMMs). The dependent variable to be modeled is a binary variable indicating, for each frame/lexeme, whether the participant selected the AccPl form that was defined as the "correct" AccPl form of that lexeme. Consequently, analyses were performed using logistic regression as implemented by the `glmer` function with the argument `family="binomial"`. Assuming that any information about GenSg and NomPl forms used by participants will potentially shift their judgments toward the AccPl form(s) that correspond to those GenSg and NomPl suffixes, this measure makes it possible to assess whether participants are using information in the provided GenSg and NomPl forms. In short, this approach allows us to estimate whether and how much knowing the GenSg or NomPl of a nonce word (or both) improves participants' likelihood to select its proper AccPl form, and therefore whether participants make use of more than a single base.

The fixed effects under investigation—in other words, the predictors or experimentally manipulated independent variables—are the effects of knowing a lexeme's GenSg or NomPl form. Each of these is a binary variable (i.e. a variable that takes either the value 1 or the value 0) indicating whether or not that form was included in the lexeme's presentation frame. In fitting a linear model, each fixed effect receives an estimate of its *coefficient*, that is, a real number which informally serves the purpose of indicating how strong an effect that predictor exhibits given the data. These variables `knew.gensg` and `knew.nompl` each evaluate to 0 when the GenSg or NomPl (respectively) was not presented in a trial, and to 1 when that form was presented. The interaction between these two variables was also examined as a fixed effect, the variable `knew.nom:knew.gen` which

evaluates to 1 only when both the GenSg and NomPl forms were presented. An intercept, which always evaluates to 1, is also included to express the baseline performance of participants when shown only the DatPl form. The log odds of selecting a lexeme's proper AccPl form under this model therefore corresponds to the sum of the coefficients of variables that evaluate to 1. For example, it predicts the success rate of a participant shown only the DatPl of a lexeme to be based on just the intercept's coefficient, while it predicts the success rate when all three bases are provided to be based on the sum of the intercept's coefficient, the coefficient of knowing the GenSg, the coefficient of knowing the NomPl, and the coefficient of knowing both the GenSg and NomPl (i.e. their interaction).

In GLMMs, random effects are included to account for variability within populations that have been sampled from. In the case of this experiment, participants have been sampled from the population of all Icelandic speakers, and stems have been sampled from the population of all possible Icelandic nonce stems. Whether any particular Icelandic speaker or stem behaves or is treated in some particular way is not within the scope of the research questions, but by adding random effects for these variables, it is possible to improve models' ability to correctly assess the effects of fixed effect variables. Random effect structures up to the maximal structure were considered, specifically random intercepts by subject and stem and random slopes for both for all three fixed effects, as recommended by Barr *et al.* (2013). However, because models including random slopes for the two predictor variables and their interaction failed to converge, I followed the recommendation of Bates *et al.* (2015a) by using only random intercepts for the two random effects, participants and items. This model suffered from no convergence problems.

Linear models: hypothesis testing

Part (a) of the probabilistic single surface base hypothesis states that in any inflectional inference task, only information from one base form can affect judgments about the distribution over potential derivative forms. Part (b) of the probabilistic single surface base hypothesis states that there is exactly one privileged cell in an inflectional system whose forms can be used as the basis of inferences about derived forms in other cells. In the context of a GLMM analysis of these experimental data, part (b) predicts that only one (or neither) variable out of `knew.gensg`

and `knew.nompl` should have a significant effect, and part (a) predicts that even if both have a significant effect, their interaction `knew.gensg:knew.nompl` should have a significant *negative* coefficient to offset the main effects' sum of coefficients from yielding a value higher than either one yields alone. The results of fitting a GLMM with the parameters described above to the experimental data, as shown in Figure 3.15, are inconsistent with these predictions.

	Dependent variable:
	correct
knew.nom	0.738^{***}
	(0.120)
knew.gen	0.423^{***}
	(0.121)
knew.nom:knew.gen	0.069
	(0.166)
(Intercept)	-1.585^{***}
	(0.303)
Observations	3,934
Log Likelihood	−1,864.399
Akaike Inf. Crit.	3,740.798
Bayesian Inf. Crit.	3,778.463
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 3.15: A GLMM with maximum likelihood coefficients predicting whether a participant's selected AccPl corresponded to the correct AccPl. Values listed for each predictor are their coefficient estimates, and associated values in parentheses are their standard errors. Table created using Stargazer (Hlavac, 2013).

Because the model uses logistic regression, coefficient estimates are log odds. The intercept estimate of -1.585 shown here, taking also into account the random effects, corresponds to "success" rate of approximately 17% on trials when only the

DatPl was presented. Knowing the GenSg (`knew.gen`) and NomPl (`knew.nom`) here both have significantly positive effects, yielding a predicted success rate of approximately 24.7% when the GenSg is known and 29.9% when the NomPl is known. The predicted success rate given both informative bases was 41.2%. These predicted success rates were calculated by exponentiating sums of model coefficients to arrive at odds and then converting these odds to probabilities.

Recall that hypothesis part (b) depends crucially on the truth of hypothesis part (a). By disproving part (a), then, it is possible to disprove the entire probabilistic single surface base hypothesis. Indeed, the fitted model parameters falsify part (a), the stipulation that a speaker can make use of information from only one base form. Note that as a linear model, the GLMM shown in Figure 3.15 is linearly additive: the models's predictions are based on the sum of its predictors' effects. Consequently, because both `knew.gen` and `knew.nom` have significantly positive effects and no significant negative interaction effect, it is possible to conclude that when participants are provided both the GenSg and the NomPl, they achieve a higher rate of correct responses than when provided only one base or the other, adding their GenSg and NomPl coefficients together. In addition, these results are incompatible with the hypothesis that while each Icelandic speaker can make use of only one privileged base, some speakers use the GenSg while others use the NomPl. A significantly negative interaction effect would indicate that both GenSg and NomPl knowledge contribute significantly but participants with knowledge of both perform less well than would be expected from their joint knowledge; however, the interaction effect achieves a *p-value* of $\approx$0.60, and the estimated coefficient is dwarfed by those of the other effects. The fact that there is no significant interaction effect indicates that the additive nature of the `knew.gensg` and `knew.nompl` effects is legitimate. Therefore the judgments of even individual Icelandic speakers can be affected by both base forms together.

Fixed effect	χ^2	p	degrees of freedom
`knew.gen`	16.94	< 0.001	2
`knew.nom`	17.82	< 0.001	2
`1` (only intercept)	110.1	< 0.0001	1

Figure 3.16: Results of likelihood ratio tests performed using the `anova` function in R (R Core Team, 2013) comparing the superset model shown in figure 3.15 to three subset models.

It is also possible to explicitly test part (b) of the hypothesis. If there is only a single privileged base form, the GenSg or NomPl (or some other form), then we would expect a simpler GLMM with a fixed effect only for knowledge of the GenSg or for knowledge of the NomPl (or neither) to achieve as much predictive power as the model shown above with a superset of these predictors. Such simpler models were fitted to the experimental data, and likelihood ratio tests were performed to perform model comparison. Figure 3.16 shows the results of these tests. The superset model predicts the experimental data better than any of the subset models under consideration, lending further evidence that there is no single privileged base form available for Icelandic speakers to use in inflectional inference, but rather that they make use of information from multiple bases.

Because participants were recruited largely through Icelandic social media forums and networks with large proportions of university-educated people, one question in the post-study demographic questionnaire asked participants whether they had taken any Icelandic language or linguistics classes at the post-secondary level. Such classes sometimes explicitly discuss the inflectional system of Icelandic (Gunnar Ó Hansson, p.c.), potentially causing participants' judgments to be informed in large part by conscious, explicit knowledge of grammatical patterns and prescriptive norms as opposed to only their intuitions as native speakers. Approximately one third of participants indicated that they had taken at least one of these advanced Icelandic language or linguistics classes (n=45 out of 123). As confirmed by adding a fixed effect encoding this binary difference among participants to the superset model, the significant main effects and lack of a negative interaction ef-

fect shown in figure 3.15 still hold of both groups of participants. However, the main effect coefficient estimates are higher for the 45 participants who had taken an advanced Icelandic class, as might be expected if they have access to explicit prescriptive knowledge in addition to their implicit native speaker intuitions. Even so, these data suggest that the GLMM analysis results do not stem solely from participants' knowledge of prescriptive linguistic norms.

3.3.3 Discussion of Icelandic experiment

The results from this experiment on native Icelandic speakers force a sound rejection of the probabilistic single surface base hypothesis as defined in 3.1. Each base form provided to participants improved their success rate, indicating that both bases can be used in inference, even in conjunction with each other. Despite the formal, empirical, and learnability-theoretic advantages of models of inflectional morphology in which information from multiple bases cannot be combined, actual speakers do not behave so simply.

Despite the falsification of the single surface base hypothesis, it is not necessarily the case that information from any and all base forms can be combined freely. With an eye toward formulating as parsimonious a theory as is compatible with available behavioral data, the next section describes a successor to the single surface base hypothesis. This hypothesis shows promise in constraining the space of possible inflectional grammars while still maintaining compatibility with the Bayesian view adopted in this dissertation.

3.4 Base independence

Having established that an inflectional system is not limited to only a single privileged base cell, and also that native speakers can combine information from multiple base cells in their inferences about unknown inflected forms, I now turn to questions of the precise nature of *how* speakers combine information from multiple base forms. Specifically, I present a simplifying hypothesis of the potential interactions among information from different base forms, the *base independence hypothesis*, and I demonstrate that the Polish nominal inflection system presents a test case for this hypothesis. An experimental investigation of Polish speakers' knowledge of this part of their inflectional system, however, failed to demonstrate even the baseline behavior that was assumed for purposes of the experimental de-

sign, and so this experimental investigation of the base independence hypothesis provides no substantial evidence for or against it.

3.4.1 The base independence hypothesis

Because speakers are able to use multiple base forms in inference, the conditional probability distribution that their grammars generate is one over derivative forms (or, equivalently in sublexical morphology, sublexicons) conditional on all or at least multiple known base forms, i.e. $p(derivative|base_1, base_2, ...base_n)$. Through application of Bayes's theorem, it can be seen that this distribution is proportional to $p(base_1, base_2, ...base_n|derivative)p(derivative)$, i.e. the joint probability distribution over all possible forms of the available bases conditioned on the derivative candidates, times the prior probability distribution of the derivative candidates. However, calculating a joint probability distribution over the forms in all base cells together may pose considerable difficulties, because it may require not only the calculation of form probabilities for all base cells individually, but also the calculation of form probabilities for all *combinations* of base cells.

For the purpose of determining whether the full joint probability distribution must indeed be calculated, I propose the *base independence hypothesis*, which states that for any lexeme in an inflectional system, any proper subset of its forms are conditionally independent given some other form. Less formally, this hypothesis states that calculating the probabilities of individual base forms suffices to arrive at the probability of those base forms together, because these individual probabilities can simply be multiplied together. For the purposes of determining a probability distribution over derivative forms given a set of base forms, it is possible to state this hypothesis mathematically by applying the definition of conditional probability, as shown in equation 3.6. Note again here that B_{cell} indicates the set of a lexeme's possible base forms in a particular cell, and S indicates the set of sublexicons.

$$p(B_1, B_2, ...B_n|S) = p(B_1|S)p(B_2|S)...p(B_n|S) \tag{3.6}$$

As equation 3.6 makes clear, the base independence hypothesis would allow the joint distribution of multiple bases to be calculated by simply calculating the probability of each base separately and then multiplying these probabilities. If this

hypothesis holds, then even with multiple bases available for inference, actually performing inference with multiple bases would likely still be tractable, as it would be equivalent to performing single-base inference for each available base and then combining these results afterwards.

It is worth considering at this point how to begin approaching the task of implementing the base independence hypothesis under the assumptions of sublexical morphology. The term $p(b_1, b_2, ...b_n|s)$, i.e. the joint probability of the provided bases given a particular sublexicon, is calculated by having the gatekeeper grammar of sublexicon s assign a probability to the base forms $b_1, b_2, ...b_n$. A sublexicon's gatekeeper grammar is parameterized by a set of weighted constraints. Each of these constraints is indexed to a specific base cell, e.g. a constraint *3Sg: a#* which evaluates to 1 if a provided 3Sg base form ends in [a] and evaluates to 0 otherwise. However, each sublexicon has only one gatekeeper grammar, which can include constraints that refer to any and all cells in the inflectional system: mixed in with *3Sg: a#* might be constraints like *3Pl: an#* and *1Pl: amos#*. Consequently, a gatekeeper grammar evaluates the probability of all provided base forms together, yielding their joint probability.

However, if the base independence hypothesis is true, then there would be a different possible mechanism for calculating the joint probability of bases. In this case, a sublexicon could have one gatekeeper grammar for each base cell, with each grammar only containing constraints specific to its cell. To arrive at the joint probability of a set of bases, then, one could have each grammar evaluate the probability of its respective base and then multiply these probabilities together. Such a set-up may seem to needlessly complicate the machinery of sublexicons, but by investigating where the predictions of these two configurations differ, it can be shown that in fact this latter set-up tightly restricts the space of possible grammars.

Predictions of the base independence hypothesis

In attempting to falsify the base independence hypothesis, it is useful to consider what sorts of patterns a model of grammar abiding by that hypothesis would be fundamentally incapable of predicting. One such type of pattern would require *cross-base constraint conjunctions*, i.e. constraints in the gatekeeper grammar which conjoin conditions on different bases. One example might be a constraint that evaluates

to 1 if and only if the provided NomSg form ends in [-o] **and** the provided GenSg form ends in [-i]. If the base independence hypothesis is implemented using a gatekeeper grammar that has a different "sub-grammar" for each base cell, and which then simply multiplies these sub-grammars' probabilities of the provided base forms, then there is no way that a constraint referring to multiple base forms could be included in such a model, as it would be incapable of ever assessing a violation.

Such constraints are not necessarily required for a grammar to express the implicational relationships in Icelandic noun inflection covered in the previous sections. For the Icelandic exponents under discussion, knowledge of a lexeme's GenSg form and its NomPl form can combine additively, as shown in the linear model in figure 3.15: there is no significant interaction term, either positive or negative, and so participants did indeed exhibit this purely additive behavior. In gatekeeper grammars, which are formally similar to logistic regression models in that they allow linear combinations of weights, the same logic would apply. Multiplying probabilities of independently evaluated bases would be equivalent to summing their weights together within a single grammar, and so the single-gatekeeper and one-gatekeeper-per-cell configurations would not differ substantially in their predictions.

But what sort of inflectional system might require such a cross-base constraint conjunction? Consider the toy system in Figure 3.17, which shows the suffixes characteristic of the 1st, 2nd, and 3rd person forms of verbs in three conjugation classes of a hypothetical language. The crucial aspect of this dataset is that each of the cells has only two possible suffixes, but the cell whose suffix differs from the other two cells' suffixes varies across the three inflectional classes.

	1Sg	2Sg	3Sg
Class 1	**-a**	**-i**	**-an**
Class 2	-a	-e	-en
Class 3	-o	-i	-en

Figure 3.17: A schematic of an inflectional system which would be able to make use of cross-base constraint conjunctions. Suffixes shown in boldface are those referred to in the hypothetical inference task below.

To see why this dataset might require a cross-base constraint conjunction, we can suppose that a speaker is attempting to predict the 3Sg form of a lexeme from its 1Sg and 2Sg forms. The 1Sg and 2Sg forms known to the speaker take the suffixes *-a* and *-i*, respectively. Let us suppose also that the three classes are equally frequent in the lexicon. To an analyst, it would be clear given the inflectional system that this lexeme should belong to Class 1 and therefore should take the 3Sg suffix *-an*, and one might predict that native speakers would share a strong judgment to this effect.

However, applying the base independence hypothesis to this scenario predicts less certainty on the part of the speaker. Equations 3.7 and 3.8 below show calculations of each 3Sg suffix candidate's pre-normalization conditional probability conditioned on the provided 1Sg and 2Sg base forms. These calculations even assume conservatively that the grammars generating conditional probabilities produce nearly categorical judgments (probabilities of 1.0, 0.5, or 0.0) by mirroring the implicational relationships seen in Figure 3.17.

$$
\begin{aligned}
& p(3Sg = an | 1Sg = a, 2Sg = i) \\
& \propto p(1Sg = a, 2Sg = i | 3Sg = an) p(3Sg = an) \text{ [Bayes's theorem]} \\
& = p(1Sg = a | 3Sg = an) p(2Sg = i | 3Sg = an) p(3Sg = an) \text{ [BIH]} \\
& = 1.0 * 1.0 * 0.\overline{3} \\
& = 0.\overline{3}
\end{aligned}
\tag{3.7}
$$

$$
\begin{aligned}
&p(3Sg = en | 1Sg = a, 2Sg = i) \\
&\propto p(1Sg = a, 2Sg = i | 3Sg = en) p(3Sg = en) \text{ [Bayes's theorem]} \\
&= p(1Sg = a | 3Sg = en) p(2Sg = i | 3Sg = en) p(3Sg = en) \text{ [BIH]} \\
&= 0.5 * 0.5 * 0.\overline{6} \\
&= 0.1\overline{6}
\end{aligned}
\tag{3.8}
$$

These equations make it clear that a 3Sg *-an* suffix is more probable given the provided base forms if assuming the base independence hypothesis, but the difference in the probability assigned to a 3Sg *-an* suffix in equation 3.7 and that assigned to a 3Sg *-en* suffix in equation 3.8 is not substantial enough to recapitulate the analyst's nearly categorical judgment that the 3Sg suffix in this case should be *-an*. In fact, as shown in equation 3.9, the probability assigned to a 3Sg *-an* suffix by a Bayesian model assuming the base independence hypothesis is only $0.\overline{6}$. Note here that Z is a normalization constant which ensures that the probabilities of the 3Sg candidates form a proper probability distribution by summing to 1; Z is equal to the sum of the values calculated in equations 3.7 and 3.8, which themselves are not probabilities because they do not sum to 1.

$$
\begin{aligned}
&p(3Sg = an | 1Sg = a, 2Sg = i) \\
&= p(1Sg = a, 2Sg = i | 3Sg = en) p(3Sg = an)/Z \text{ [Bayes's theorem]} \\
&= 0.\overline{3}/(0.\overline{3} + 0.1\overline{6}) \\
&= 0.\overline{6}
\end{aligned}
\tag{3.9}
$$

However, this mismatch between intuitive and predicted probability distributions is crucially due to application of the base independence hypothesis, not due only to the use of Bayes's theorem. (Note also that this model is not recapitulating the lexical frequencies of 3Sg exponents: it is *-an* that receives a probability of $0.\overline{6}$, not *-en*.) Equations 3.10 and 3.11 show the pre-normalization conditional probabilities of 3Sg *-an* and *-en*, respectively, *without* the base independence hypothesis but otherwise under the same assumptions. The probability of 3Sg taking the suffix *-en* in this case is predicted to be zero, meaning that the probability of the suffix *-an*

is 1.0, matching the intuition of an analyst applying knowledge of the inflectional system in Figure 3.17 to the inference task. Equation 3.12 shows this calculation.

$$\begin{aligned}
&p(3Sg = an | 1Sg = a, 2Sg = i) \\
&\propto p(1Sg = a, 2Sg = i | 3Sg = an) p(3Sg = an) \text{ [Bayes's theorem]} \\
&= 1.0 * 0.\overline{3} \\
&= 0.\overline{3}
\end{aligned} \tag{3.10}$$

$$\begin{aligned}
&p(3Sg = en | 1Sg = a, 2Sg = i) \\
&\propto p(1Sg = a, 2Sg = i | 3Sg = en) p(3Sg = en) \text{ [Bayes's theorem]} \\
&= 0.0 * 0.\overline{6} \\
&= 0.0
\end{aligned} \tag{3.11}$$

$$\begin{aligned}
&p(3Sg = an | 1Sg = a, 2Sg = i) \\
&= p(1Sg = a, 2Sg = i | 3Sg = en) p(3Sg = en) / Z \text{ [Bayes's theorem]} \\
&= 0.\overline{3} / (0.\overline{3} + 0.0) \\
&= 1.0
\end{aligned} \tag{3.12}$$

We can now consider the distinction between these two scenarios, one with the base independence hypothesis and one without, in the constraint-based terms of MaxEnt harmonic grammars. If constraints are restricted to referring to only one base each, then the relevant constraints would be [1Sg: -a] and [2Sg: -i]. In these constraints, the portion to the left of the colon indicates the base cell whose form is evaluated, and the portion to the right of the colon indicates the material whose presence in the selected base cell's form results in the constraint evaluating to 1. With access only to these two constraints (but not their conjunction), the grammar would be limited to evaluating the violation profiles of the 1Sg and 2Sg forms separately, tantamount to evaluating their conditional probabilities independently.[2]

[2]Such a constraint set would make the grammar functionally equivalent to the naïve Bayes formalism used in statistics, information retrieval, and machine learning (Lewis, 1998).

Accordingly, the grammar would fail to assign these bases a near-zero joint probability when conditioned on the sublexicon corresponding to the 3Sg taking the *-en* suffix.

The cross-base constraint conjunction which would be useful for such a system is [1Sg: -a & 2Sg: -i], which evaluates to 1 only if both its requirements are met. With access to this feature, the MaxEnt grammars would be able to assign the corresponding base forms conditional probabilities much closer to those shown in equations 3.10 and 3.11, paralleling the analyst's intuition of a (nearly) categorical prediction.

The constraint-based interpretation of the base independence hypothesis, in which the hypothesis explicitly forbids cross-base constraint conjunctions, makes the learnability-theoretic appeal of the base independence hypothesis clear. The search space for phonological constraints within a single set of forms is formidably large (Hayes & Wilson, 2008). If the space of possible constraints also includes cross-base constraint conjunctions, then the size of this search space is multiplied by the number of base cells whose constraints could conceivably be conjoined; at the very least, assuming that constraint conjunctions maximally conjoin constraints on two bases, this allowance expands the search space from n constraints to n^2 constraints. The following subsections describe an experiment intended to test whether cross-base constraint conjunctions are indeed necessary, designed in the hopes of demonstrating that linguists need not worry about this potential explosion of the constraint search space.

3.4.2 Testing the base independence hypothesis

The previous subsection has described a hypothetical inflectional system for which a Bayesian model of morphology adhering to the base independence hypothesis would make substantially different predictions from a similar model without the base independence hypothesis. As with the single surface base hypothesis, the next questions are whether such patterns exist in the inflectional systems of natural languages, and whether speakers' judgments are consistent with the hypothesis's predictions.

The first question, at least, I can answer with a confident "yes": this subsection describes a slice of the Polish nominal inflectional system which qualitatively

parallels the toy example in Figure 3.17. I motivate and describe an experiment on native speakers of Polish designed to test the base independence hypothesis on the basis of this pattern, largely following in the methodological footsteps of the Icelandic experiment. Ultimately, the experiment fails to falsify the base independence hypothesis, but more investigation is necessary before concluding that the base independence hypothesis holds true in general, especially given some unexpected behavior patterns among participants.

Polish soft declensions

Nouns in Polish inflect for case and number, with seven cases and two numbers (Schenker, 1955). Of particular present interest are the nominative and genitive forms, whose abbreviations in the singular and plural are shown in Figure 3.18.

	Nominative	Genitive
Singular	*NomSg*	*GenSg*
Plural	*NomPl*	*GenPl*

Figure 3.18: The two cases of immediate interest and two numbers of Polish nouns, as well as their abbreviations. The other cases are accusative, locative, dative, instrumental, and vocative.

Like Icelandic, nouns in Polish carry information about grammatical gender, and inflectional classes (as defined by constellations of suffixes) and gender interact in complex but predictable ways. Each of the three genders—masculine, feminine, and neuter—corresponds to multiple inflectional classes, but one commonality is that each gender minimally corresponds to a *hard* class and a *soft* class of nouns, with this distinction based on whether the final consonant of the lexeme stems of a class is phonologically "hard" (typically non-palatalized) or "soft" (typically palatalized).

	Hard feminine		Soft feminine	
	Singular	Plural	Singular	Plural
Nominative	map**a**	map**y**	granic**a**	granic**e**
Accusative	map**ę**	map**y**	granic**ę**	granic**e**
Genitive	map**y**	map	granic**y**	granic
Locative	mapi**e**	map**ach**	granic**y**	granic**ach**
Dative	mapi**e**	map**om**	granic**y**	granic**om**
Instrumental	map**ą**	map**ami**	granic**ą**	granic**ami**
Vocative	map**o**	map**y**	granic**o**	granic**e**

Figure 3.19: The full inflectional paradigms of nouns MAP *map-* and BORDER, LIMIT *granic-*, representative of the hard feminine and soft feminine inflectional classes, respectively. Paradigms are shown orthographically, with suffixes given in bold. Forms other than the GenSg, GenPl, NomPl, and DatPl are shown here only to give an overall sense of the inflectional system.

For purposes of testing the base independence hypothesis, I focus now on the GenSg, GenPl, and NomPl forms of the three genders' soft declensions, as these three stand in a relationship equivalent to the hypothetical system sketched previously. Figure 3.20 shows the suffixes associated with each of these cells in the three classes. Note that the patterns of similarity and difference here perfectly parallel those shown in the toy dataset in Figure 3.17. GenSg, GenPl, and NomPl take the place of 1Sg, 2Sg, and 3Sg, and the 3Sg exponents *-a* and *-e* take the place of the hypothetical (but coincidentally similar) *-an* and *-en*. The two exponents each of the GenSg and GenPl are distributed with respect to NomPl exponents in the same way as those the 1Sg and 2Sg in figure 3.17 were distributed with respect to the 3Sg.

	GenSg	GenPl	NomPl
Soft neut.	-a	∅	-a
Soft masc.	-a	-y	-e
Soft fem.	-y	∅	-e

Figure 3.20: The suffixes associated with the GenSg, GenPl, and NomPl forms of soft neuter, masculine, and feminine nouns in Polish.

Because this portion of the Polish nominal inflectional system exhibits the same properties of inter-predictiveness as the toy dataset in Figure 3.17, Polish shows potential as a testing ground for the base independence hypothesis. The task of predicting the NomPl of a lexeme from knowledge that its GenSg takes the suffix *-a* and its GenPl has a null suffix (i.e. when the lexeme belongs to the soft neuter class) serves as the key test case. The rest of this subsection describes an experiment designed to test whether Polish speakers in this situation prefer the *-a* suffix for the lexeme's NomPl more often than would be expected under the base independence hypothesis.

Methodology

This experiment was carried out using a methodology very similar to that used for the Icelandic experiment described in 3.3.1. A total of 219 Polish native speaker participants were recruited by posts on mailing lists and through word of mouth. The experiment itself was carried out entirely online using the Experigen (Becker & Levine, 2012) framework.

On the Experigen web interface, participants were given information about the experiment and were then provided a consent form to electronically sign. Each participant who consented engaged next in two non-randomized practice trials, after which she or he completed forty-eight test trials, one for each lexeme. Finally, all consenting participants filled out a demographic questionnaire asking for non-identifying personal information. All parts of the experiment were presented in Polish, as translated from English by a native speaker of Polish. The English translation of this questionnaire can be found in the appendix.

Each trial concerned a single novel lexeme designed to resemble existing Polish

noun lexemes. A participant would first be exposed to some number of inflected forms of the novel lexeme in carrier sentences. The task was then to select one preferred NomPl form of the lexeme out of a pair of options, one taking an *-a* suffix and the other taking an *-e* suffix. The order of these response options was randomized. The participant's choice of NomPl form was recorded as the key experimental measure.

The information about a novel lexeme presented to a participant before the NomPl choice task varied according to each stimulus frame's *presentation condition*. This variable ranged across the four options shown below, which correspond to the inflected forms shown to participants before they were asked to select an NomPl form. The DatPl form, which conveys no information about which NomPl suffix is most appropriate, is always provided to introduce participants to each novel lexeme and encourage them to think of it as an existing Polish noun.

1. DatPl [uninformative]
2. DatPl, then GenSg [somewhat informative]
3. DatPl, then GenPl [somewhat informative]
4. DatPl, then GenSg, then GenPl [maximally informative]

Figure 3.21: The four presentation conditions of the Polish experiment.

Novel lexeme stems, 48 in total, were assigned randomly to the three inflectional classes: soft masculine, soft feminine, and soft neuter. These stems were designed so as to minimally influence judgments about their appropriate inflectional class or gender. Specifically, all stems were of the shape CVCVC, with their last vowels and final consonants drawn from the sets {i, y, o, a} and {ń, ś, ź, ć, }, respectively. This set of final consonants was chosen in collaboration with a natively Polish-speaking linguist so as to include a phonologically diverse (yet small) set of unambiguously soft consonants, and similarly the set of vowels was primarily chosen so as to eliminate concerns about potential stem changes, e.g. *yer* deletion (Jarosz, 2005; Scheer, 2012). Three stems were generated from each com-

bination of vowel and final consonant, yielding a total of 48 stems. The three stems with the same vowel-consonant combination were always assigned to different inflectional classes. To minimize the risk of stems being similar enough to existing stems that the existing lexemes' inflection could unduly affect judgments about the novel stems, I implemented a script that rejected any stems with an edit distance of 1 from any existing stem in a publicly available one-million word subcorpus of the *National Corpus of Polish* (Przepiórkowski *et al.*, 2010). The aforementioned Polish-speaking linguist also performed a similar form of filtering using her own judgments of similarity. A complete list of stimuli can be found in the appendix.

(Pseudo-)randomization of stimuli was performed in two ways. The order of novel lexemes was first itself randomized. The assignment of lexemes to presentation conditions also varied across participants, but rather than being randomized, Experigen cycled through three possible sets of pairings of lexeme and presentation condition from each participant to the next. Each of of these sets of pairings evenly distributed lexemes of various inflectional classes evenly across the three presentation conditions, such that there were always exactly twelve lexemes in each inflectional class in each presentation condition, one for each consonant-vowel combination.

The DatPl, GenSg, and GenPl forms, when presented, were always given inside a carrier sentence designed to make it clear which paradigm cell the presented form belongs to. These sentences were carefully designed so as not to provide any additional information about gender. The NomPl was also elicited using a carrier sentence, which provided an underscore in the position of the requested NomPl form and which also provided no extra information about gender. After the presentation of each inflected form, participants pressed a button to continue on to the next inflected form or finally to the NomPl choice task. This button-based procedure was added to encourage participants to consider the information provided by each inflected form. Inflected forms themselves were bolded to make them stand out against the carrier sentences.

Together, all of the carrier sentences formed a short narrative about children's toys, intended to allow participants to think of novel lexemes as the names of obscure Polish toys. Toys were chosen as the topic because toy words in Polish are not restricted to having specific genders, as are e.g. animal names. The narrative

was coherent regardless of which sentences were included. The purpose of this design choice, in addition with the explicit instruction before the test trials to think of novel words as real but rare Polish nouns, was to encourage participants to use their knowledge of inflectional patterns in existing Polish words when performing the NomPl choice task. All presentation and elicitation was performed using orthographic representations, and no recordings were played or made at any time. Frame sentences are provided in Polish with English translations in the appendix.

Results

The base independence hypothesis predicts that participants given all three base forms will perform no better than predicted by the summed effects of their gains in accuracy attributable to knowing the GenSg form or the GenPl form separately. Indeed, as the visual summary of responses in figure 3.22 sketches, success rates in trials including all three base forms were no higher than those presenting only two base forms. Unexpectedly, however, there were no measurable differences in accuracy at all across *any* of the four presentation conditions.

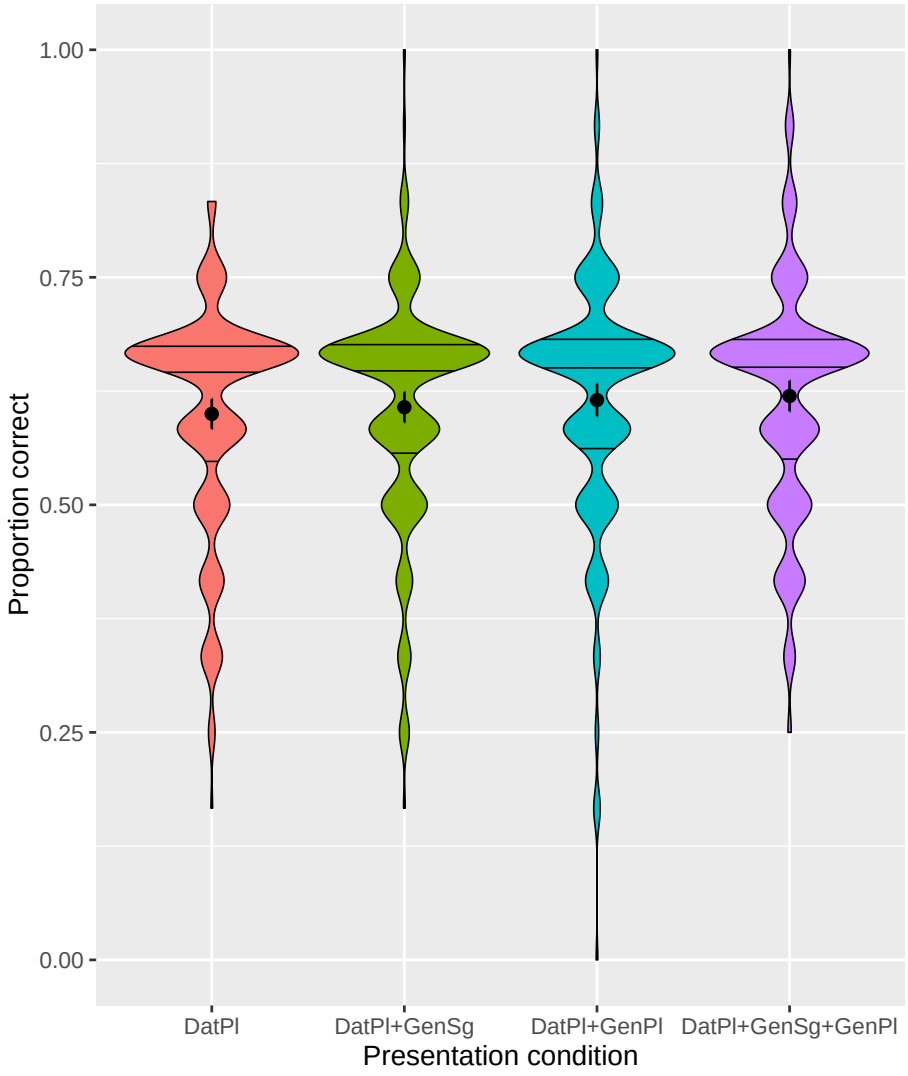

Figure 3.22: Participants' proportions of "correct" responses in the Polish experiment by presentation condition. Vertical bars show 95% confidence intervals, and horizontal bars show quartile values. Color-coded probability density functions show kernel estimations of the underlying distributions.

These results do *not* indicate that participants failed to ever make use of information contained in presented base forms. Rather, the apparent lack of any differences here is due to considerable discrepancies in behavior when presented lexemes belonging to the three gender classes (even though participants needed to infer each lexeme's gender). Recall the organization of the key exponents in Polish, as repeated in figure 3.23.

	GenSg	GenPl	NomPl
Soft neut.	-a	∅	-a
Soft masc.	-a	-y	-e
Soft fem.	-y	∅	-e

Figure 3.23: The suffixes associated with the GenSg, GenPl, and NomPl forms of soft neuter, masculine, and feminine nouns in Polish.

While the base independence hypothesis—if valid—should hold true for the entirety of the inflectional system, it is only the neuter lexemes that serve as the clearest test case for the hypothesis. This is because only among neuters (out of the three classes shown here) does neither the GenSg exponent nor the GenPl exponent uniquely identify a lexeme's gender: knowing that a lexeme's GenSg ends in *-y* suffices to categorize a lexeme as feminine, while knowing that a lexeme's GenPl ends in *-y* suffices to categorize a lexeme as masculine. The appropriate question to ask, then, is whether participants are more likely to select an *-a* NomPl form when provided both base forms than would be expected by simply combining the effects of knowing the GenSg or GenPl form individually. In other words, is the amount by which participants improve from the baseline (only DatPl) when given both the GenSg and GenPl attributable purely to the simple combination of improvements seen when provided each one of these forms separately? Figure 3.24 shows rates of selecting the correct NomPl form (*-a*) among neuter lexemes, aggregated by presentation condition.

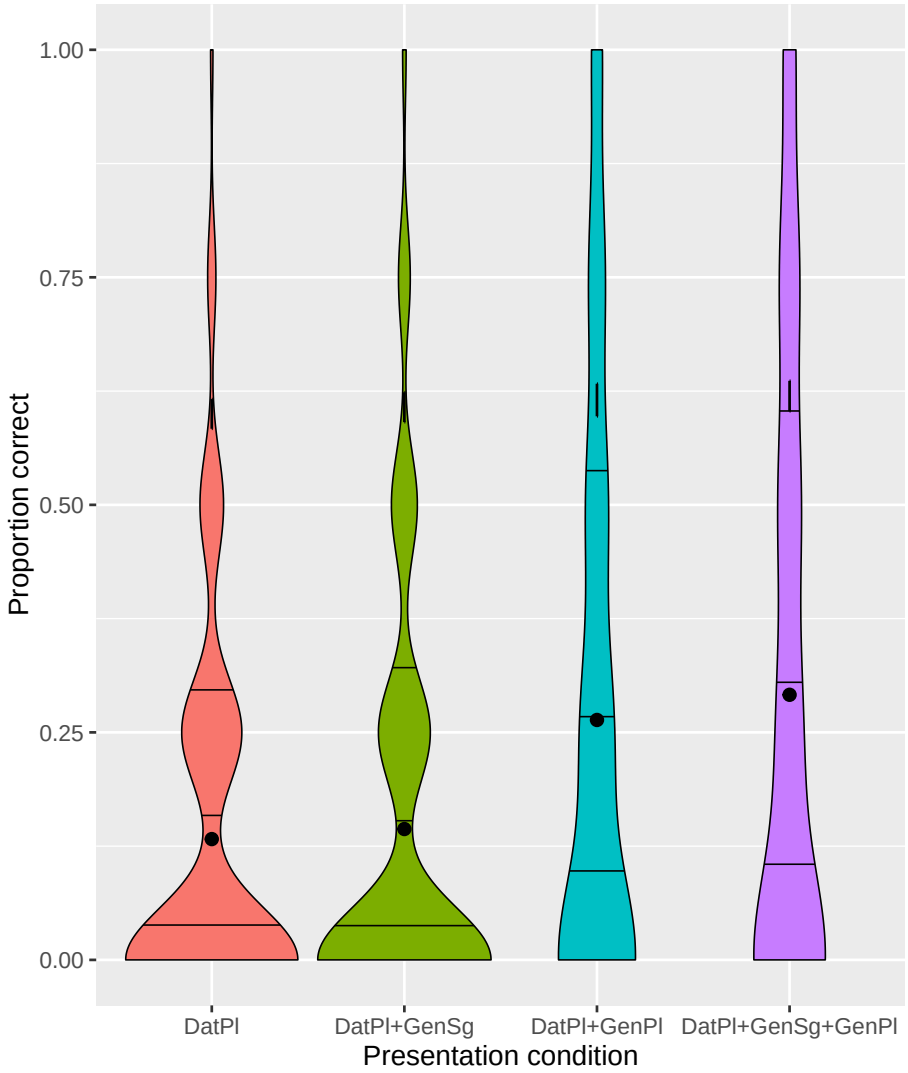

Figure 3.24: Participants' proportions of "correct" (*-a* NomPl) responses for neuter-class items in the Polish experiment by presentation condition. Vertical bars show 95% confidence intervals, and horizontal bars show quartile values. Color-coded probability density functions show kernel estimations of the underlying distributions.

Even when looking at just the neuter lexemes, participants fail to ever demonstrate highly accurate "superadditive" behavior when provided both the GenSg and GenPl, a result consistent with the base independence hypothesis. Even so, participants' behavior within gender classes also defied many expectations, just as did their behavior in the aggregate as shown in figure 3.22, and so I conclude that it is premature to take these results as clearly validating the base independence hypothesis.[3] The remainder of this section serves to motivate these claims statistically and explore possible explanations of these results.

In order to provide statistical backing to these claims and to further formalize and test them, I analyzed the results of this experiment in R (R Core Team, 2013) using generalized linear mixed effect models (GLMMs) as implemented by the `glmer` function in the `lme4` package (Bates *et al.*, 2015b), just as I did for the results of the Icelandic experiment described in 3.3.2. Again in parallel to the Icelandic experiment, the dependent variable is a binary variable indicating whether the participant's chosen NomPl form is the same as the lexeme's intended ("correct") NomPl form. Consequently, the model used logistic regression by setting the `family` parameter to `binomial`. Fixed effects are included for binary variables indicating whether the GenSg form was presented (`knew.gensg`) and whether the GenPl form was presented (`knew.genpl`), as well as their interaction (`knew.gensg:knew.genpl`). By-participant and by-lexeme random intercepts are also included, but random slopes were excluded due to model convergence issues; see 3.3.2 for details of the motivations behind these choices.

GLMMs are a natural fit to the research question addressed here. With knowledge of only the DatPl as the baseline (corresponding to the intercept), the effects of `knew.gensg` and `knew.genpl` correspond to the increase in likelihood of selecting the correct NomPl gained by knowing the GenSg form and the GenPl form separately, that is, independent of each other. To falsify the base independence hypothesis, Polish speakers' judgments when provided joint knowledge of

[3] Of course, because the base independence hypothesis serves as the null hypothesis in this experiment, strictly speaking no result would be able to *prove* base independence. It may be possible to reframe the research question so as to cast base independence as a falsifiable alternative hypothesis with "base dependence" as a falsifiable null hypothesis, but given that these experimental results lack clear interpretations, I content myself with demonstrating that participant behavior is not incompatible with the base independence hypothesis as formulated.

the GenSg and GenPl forms would need to differ substantially from those predicted by combining effects of `knew.gensg` and `knew.genpl`. In other words, to falsify the base independence hypothesis, speakers would need to demonstrate a superadditive effect between knowledge of the GenSg and GenPl, which in a GLMM would correspond to a significantly positive interaction term for the two predictors, the variable `knew.gensg:knew.genpl`.

Figure 3.25 shows the results of fitting such a GLMM to the Polish experimental data. Not only is there no significant interaction between the fixed effects, but the fixed effects themselves do not achieve significance. The p-values of `knew.gensg`, `knew.genpl`, and their interaction are 0.6061, 0.151, and 0.8142, respectively. The effect sizes are also small compared, for example, to those from the Icelandic experiment shown in Figure 3.15.

	Dependent variable:
	correct
knew.gensg	0.036
	(0.069)
knew.genpl	0.099
	(0.069)
knew.gensg:knew.genpl	−0.023
	(0.098)
(Intercept)	0.500**
	(0.214)
Observations	10,841
Log Likelihood	−5,326.419
Akaike Inf. Crit.	10,664.840
Bayesian Inf. Crit.	10,708.580
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 3.25: A GLMM with maximum likelihood coefficients predicting whether a participant's selected NomPl corresponded to the correct NomPl. Values listed for each predictor are their coefficient estimates, and associated values in parentheses are their standard errors. Table created using Stargazer (Hlavac, 2013).

Figure 3.26 shows similar linear models fit to data only for each of the three gender classes. Notably, the dependent variable in these models is not whether a participant selected the "correct" NomPl exponent, but rather simply *which* NomPl

exponent was chosen. This variable `guessed.suffix` evaluates to 1 when a participant chooses *-e* and to 0 when a participant chooses *-a*, meaning that more positive coefficients indicate a greater propensity to select *e*-final NomPl forms rather than *a*-final forms. The reason for this change is that each gender class has a specific NomPl exponent that is "correct", and so the two dependent variable schemes are mathematically equivalent to each other (except for signs on coefficients), but using `guessed.suffix` makes it easier to see which effects were of which sign across the three gender classes. Assuming the associations between NomPl suffix and gender class shown in figure 3.23, more positive coefficients can therefore also be thought of as indicating beliefs that a lexeme is masculine or feminine, as opposed to neuter.

	Dependent variable:		
	`guessed.suffix` == *-e*		
	neuter	*masculine*	*feminine*
`knew.gensg`	−0.127	0.360**	−0.254*
	(0.148)	(0.145)	(0.134)
`knew.genpl`	−1.092***	−0.030	−0.672***
	(0.137)	(0.137)	(0.129)
`knew.gensg:knew.genpl`	−0.032	−0.402**	0.191
	(0.190)	(0.199)	(0.180)
Intercept	2.593***	2.240***	2.276***
	(0.194)	(0.170)	(0.159)
Observations	3,614	3,613	3,614
Log Likelihood	−1,581.124	−1,396.733	−1,663.658
Akaike Inf. Crit.	3,174.248	2,805.466	3,339.316
Bayesian Inf. Crit.	3,211.403	2,842.620	3,376.471
Note:			*p<0.1; **p<0.05; ***p<0.01

Figure 3.26: GLMMs with maximum likelihood coefficients predicting, for each `subset` of the data of a particular gender class, which NomPl suffix participants selected. Values listed for each predictor are their coefficient estimates, and associated values in parentheses are their standard errors. Table created using Stargazer (Hlavac, 2013).

Looking for now only at the results for the neuter lexemes in figure 3.26, participant behavior is more in line with lexical patterns than the overall results in figure

3.25. The coefficient for `knew.genpl` is significant and negative, indicating that knowledge of a neuter lexeme's GenPl form improves participants' accuracy in selecting the *-a* NomPl exponent. However, the main effect for `knew.gensg` is not significant; it is not clear what could be causing this difference between the two effects. However, most importantly, their interaction `knew.gensg:knew.genpl` is not significant and has a small coefficient estimate. This result is consistent with the base independence hypothesis.

Note that all three gender classes' models have a significant intercept with a large, positive coefficient. These coefficients indicate that when only a lexeme's DatPl is presented—in which case participants should have no ability to discriminate among the three gender classes—there is a strong bias toward selecting the NomPl form ending in *-e*. Moreover, for neuter nouns, even when the coefficient estimates for all fixed and random effects are added up, i.e. modeling the DatPl+GenSg+GenPl condition, the sum of this negative value -1.251 and the intercept 2.593 still yields a positive value. Therefore even when presented with bases which should allow participants to always select the "correct" *-a* suffix for these forms, participants still tend to prefer selecting the *e*-final forms. Figure 3.24 from earlier in this section can be seen as a visualization of this result: note that even in the DatPl+GenSg+GenPl condition, participants still chose the "correct" *-a* suffix less than 29% of the time. As the next chapter covers in detail, this bias toward *e*-final forms can be viewed as an effect of lexical frequencies interfering with speakers' ability to make fully productive use of even categorical implicational relationships in their inflectional systems.

However, not all patterns of behavior shown by these models have such convenient explanations. As shown in the model of feminine lexeme responses, knowing their GenSg or GenPl forms actually *decreased* participant accuracy in selecting the *-e* NomPl suffix consistent with this gender class. Among masculine lexemes, knowledge of the GenSg form (but not the GenPl form) improved rates of selecting the appropriate *-e* NomPl suffix, but knowledge of *both* GenSg and GenPl resulted in *lower* accuracy than that achieved with only the GenSg. Recall that the phonological/orthographic shapes of the stems used in the experiment indicated specifically that lexemes used must belong to a soft declension, and that no other soft classes have distributions of suffixes that participants could plausibly be overap-

plying. These irregularities suggest that inexplicable confounds may have marred the design of this experiment.

As in the Icelandic experiment, one of the questions in the demographic questionnaire presented at the end of the experiment asked participants whether they had ever taken Polish language or linguistics classes since secondary school. Neither sub-population demonstrated different patterns of behavior, in general or just for neuter lexemes, from those observed for the combined population, or from each other's. Even significance levels were identical across the three sets of participants, indicating that advanced prescriptive knowledge of Polish grammar was not a likely source of any of these response patterns.

Overall, the lack of significantly positive interactions both in the response data overall and within responses for only neuter lexemes is consistent with the base independence hypothesis. Even so, participants' behavior deviated substantially in other ways from what one would expect based on the relevant lexical patterns. These aberrations suggest that the results should not be interpreted so straightforwardly as validating the base independence hypothesis, but leave open the possibility that future experiments might build on this methodology to address the question in a more conclusive way.

3.5 Summary and discussion

This chapter has presented an investigation of the ways that humans make use of known inflected base forms of a lexeme when predicting unknown derivative forms of the same lexeme. The general procedure I have taken has been to describe a simple, falsifiable baseline model of such inference and then test experimentally, for each increase in model complexity, whether native speakers' linguistic behavior justifies adding this complexity—that is, whether the behavior is consistent with the more complex model but not with the baseline.

The modeling choices under consideration in this chapter are derived from the relationships between base and derivative forms in Bayesian, surface-oriented approaches to inflectional morphology, specifically sublexical morphology. Equation 3.13 repeats the fundamental equation of sublexical morphology as an illustration of these relationships.

$$p(S|B_1, B_2, ...B_n) \propto p(B_1, B_2, ...B_n|S)p(S) \tag{3.13}$$

The first baseline tested was a probabilistic interpretation of the single surface base hypothesis (Albright, 2002), which posits that only a single base form's shape can be referred to when performing inflectional inference, and moreover that this privileged base is the same paradigm cell regardless of the nature of the inference task. In terms of equation 3.13, this hypothesis can be thought of probabilistically as defining its left-hand side as equal to $p(S|B_{privileged})$, where the cell of $B_{privileged}$ is defined as invariant across an entire inflectional system.

This chapter presented experimental evidence from an Icelandic *wug* test that this hypothesis cannot account for the inferential abilities of Icelandic speakers, and therefore that the hypothesis is not valid cross-linguistically. In the experiment, Icelandic speakers exhibited an ability not only to make use of information from both GenSg and NomPl forms when predicting AccPl forms, but also to combine information from both bases in a single inference task. I concluded, then, that the equivalence of the left-hand side of equation 3.13 to $p(S|B_{privileged})$ does not hold.

This result prompted a follow-up question: are there any principled limitations on how speakers can combine information from various available base forms? Making use of the probabilistic, Bayesian setting of sublexical morphology, I proposed one such possible limitation. The base independence hypothesis constitutes a simplifying assumption about the left term of of the right-hand side of equation 3.13, i.e. the term corresponding to the base probabilities assigned by sublexicons' MaxEnt gatekeeper grammars. According to the base independence hypothesis, probabilities of bases given a sublexicon are independent of each other, and so their joint conditional probability is simply the product of their individual conditional probabilities, as shown in equation 3.14.

$$p(B_1, B_2, ...B_n|S) = p(B_1|S)p(B_2|S)...p(B_n|S) \tag{3.14}$$

To test the base independence hypothesis, I proposed an experiment similar to that performed with Icelandic speakers, but instead targeting judgments pertaining to a relevant pattern in Polish. While the results of this experiment could be generously interpreted as a failure to falsify the hypothesis, unexpected response

patterns suggest that the experiment may not have been designed in a way that appropriately tests it.

From these empirical data, I conclude tentatively that while speakers of languages with inflectional morphology are able to combine information from multiple base forms when performing inference of unknown inflected forms, there is no conclusive evidence of hard or soft restrictions on the ways that such information can be combined. The complexity of speakers' morphological knowledge therefore represents challenges to linguists interested in modeling knowledge of inflectional morphology, both from the standpoint of creating a parsimonious theory, and from the standpoint of designing an efficient learning algorithm. Even so, I hope that others will take up the mantle of testing the base independence hypothesis and proposing other testable hypotheses about the limits of inflectional knowledge and morphological inference.

Chapter 4

Empirical priors

What does it mean for a morphological pattern to be *productive*? In investigating the implications of a Bayesian view of inflectional morphology, the present chapter addresses this fundamental question of linguistic inquiry. On the basis of experimental evidence from Icelandic and Polish, I conclude that speakers' apparent failure to apply lexically robust patterns can be fruitfully understood not as exhibiting a lack of productivity, but rather as a mathematically predictable interaction between two competing pressures. The first of these is the familiar pressure to apply the implicational relationships evidenced in the lexicon, and the second is a simpler one: a heuristic that matches lexical frequencies of morphological exponents without regard for their co-occurrence (or lack thereof) with other exponents. Tying this second pressure to the Bayesian concept of *prior probabilities*, I show that Bayesian approaches to inflectional morphology predict this interplay, and that sublexical morphology in particular provides a set of tools for understanding these phenomena not only qualitatively, but also quantitatively.

Similar to the preceding chapter, this chapter seeks to empirically validate one specific aspect of the instantiation of Bayes's theorem that forms the core of sublexical morphology. Equation 4.1 shows the key claim of a Bayesian model of inflectional morphology: that the probability of a derivative form d (among the candidates D) given a sets of known base forms B_{cell} is proportional to the likelihood of those base forms given the derivative form times the prior probability $p(D)$ of the derivative form.

$$p(D|B_1, B_2, ...B_n) \propto p(B_1, B_2, ...B_n|D)p(D) \tag{4.1}$$

The major contribution of sublexical morphology as a specific Bayesian theory of inflectional morphology is to make calculation of the likelihood term and prior probability term mathematically and intuitively straightforward. This theory sets up correspondences between derivative forms D and sublexicons S, so that equation 4.1 can be rewritten as shown in equation 4.2.

$$p(S|B_1, B_2, ...B_n) = p(B_1, B_2, ...B_n|S)p(S) \tag{4.2}$$

Crucially for this chapter, the sublexical approach entails the intuitive way to define $p(S)$ shown in equation 4.3: that the prior probability of a sublexicon is proportional to the size of that sublexicon, where size is defined as the number of lexemes associated with it. Specifically, I describe the priors used in sublexical morphology as *empirical priors*, since they are based on speakers' observations about their language.

$$p(S) \propto |S| \tag{4.3}$$

In section 4.1, I use a toy inflectional system to illustrate the differences between a traditional understanding of productivity in inflectional morphology and a Bayesian understanding that incorporates prior probabilities. Returning then to the Icelandic and Polish experiments introduced in chapter 3, sections 4.2 and 4.3 present *post hoc* analyses of the data from these experiments which demonstrate the influence of empirical prior probabilities on experimental participants' responses. These sections also address questions of how an empirical prior should be defined. A concluding section summarizes the theoretical and empirical importance of empirical priors and discusses implications.

4.1 Priors in inflection

This section describes the theoretical and empirical implications of Bayesian priors in surface-oriented inflectional morphology. First, I describe a traditional—one could say prescriptive or pedagogical—view of morphological inference. I provide a Bayesian interpretation of this conception of how such inference should proceed.

I then generalize this model to allow for *regularization*, by which priors come to play a greater role in inference, and discuss two types of priors: uniform and empirical.

Figure 4.1 shows a small, simple inflectional system for nouns in a hypothetical language. These nouns have only two forms, a singular and a plural, and each noun belongs to one of three inflectional classes. I assume for now that these three classes are equal in frequency, and that there are no nouns in the language which do not belong to one of these three classes.

	Singular	Plural
Class 1	-a	-e
Class 2	-o	-e
Class 3	-u	-i

Figure 4.1: A toy nominal inflectional system, showing the "singular" and "plural" forms of nouns in three classes.

One can now consider the implicational relationships within this system formally using the notation of probability theory. Because the singular exponents are all distinct from the plural exponents, I will use a shorthand by which, for example, $p(i|u)$ is written to mean $p(plural = i|singular = u)$, that is, the probability that the plural exponent is i given that the singular exponent is [u].

Suppose that a speaker of this language hears the singular form of an unfamiliar noun which ends in [u]. According to a prescriptive or pedagogical view of inflectional morphology, which I will call the *traditional* view, one might expect this speaker to determine conclusively that this lexeme belongs to inflectional class 3. One would then conclude that the speaker should be able to predict with perfect confidence that its plural form should end in [i]. It is also possible to come to such a conclusion without the intermediate abstraction of inflectional classes, by observing that forms whose singulars end in [u] always (in the lexicon) have plurals ending in [i]. Mathematically, this view of inflectional morphology predicts that the speaker's grammar should set $p(i|u)$ equal to 1.0. Similarly, this view would predict, for example, that $p(i|o) = 0.0$.

Such predictions can be understood in a Bayesian way. Recall that according to Bayes's theorem as applied to surface-oriented morphological inference, $p(D|B)$ is proportional to $p(B|D)p(D)$, i.e. proportional to the likelihood of the base form(s) conditioned on the derivative form times the prior probability of the derivative form. It is helpful to think of the likelihood term and the prior term as serving distinct purposes. Informally, the likelihood term evaluates how much the newly available information (in this case the base form) shifts the balance of probabilities toward or away from each outcome (derivative), while the prior term implements a bias toward pre-existing probabilities which ignores new information (like the shapes of a base form). Crucially, then, the balance between how much the overall system depends on new information like base forms rather than defaulting to prior probabilities depends on the perceived *quality* of the new information, that is, how useful the system considers that information to be. As an example, the ability of the likelihood term to shift judgments away from those based solely on prior probabilities would be greater when a speaker observes that derivative forms are easily predictable from base forms, as opposed to when a speaker considers base forms poor predictors of derivative forms.

For now, it should suffice to think in the abstract of *regularization* as a pressure which forces a new-information-sensitive likelihood term toward conservatism, in the sense of preventing it from "making strong judgments" on the basis of new information. The more regularization the system exhibits, the less the system will make use of the newly available information and the more it will rely on its prior probabilities. In practice, such regularization is a critical ingredient in creating models which are able to generalize effectively to novel data rather than "over-learning" accidental regularities in their training data. For reviews of the usefulness of regularization in statistics and machine learning, see Bickel *et al.* (2006) and Friedman *et al.* (2004), respectively, and see Wilson (2006) and Hayes (2011) for examples of the demonstrated importance of regularization in phonology.

The traditional view of inflectional morphology, by which for example $p(i|u) = 1.0$, corresponds to a complete lack of regularization. Such a model predicts exactly the conditional probabilities that it observes, and because all words with a singular [u] have a plural [i], the model makes an equivalent prediction. By preventing regularization entirely, it becomes possible to quantitatively mimic the predicted

probabilities above in a Bayesian setting, as equations 4.4 show. These equations assume for now empirical priors based on lexical frequencies: because each class has the same frequency, and because the plural [i] exponent occurs in only one class whereas the [i] exponent occurs in two, the prior probabilities of [i] and [e] are $0.\overline{3}$ and $0.\overline{6}$, respectively. This equation shows that with no regularization, the prior probabilities are rendered irrelevant by the great difference in the base likelihood terms.

$$\begin{aligned} p(i|u) &\propto p(u|i)p(i) = 1.0 * 0.\overline{3} = 0.\overline{3} \\ p(e|u) &\propto p(u|e)p(e) = 0.0 * 0.\overline{6} = 0.0 \\ \therefore p(i|u) &= 0.\overline{3}/(0.\overline{3} + 0.0) = 1.0 \end{aligned} \tag{4.4}$$

The opposite of this scenario is one with maximal regularization, which intuitively corresponds to the model ignoring any information provided by base forms. When regularization is maximized, conditional probability distributions over bases will be maximally entropic, i.e. every possible base form will be given the same conditional probability. The actual value of that probability depends only on the number of possible base forms, because they must sum to 1. In equation 4.5, maximal regularization has set both $p(u|i)$ and $p(u|e)$ equal to the same value, $0.\overline{3}$, because for either plural exponent ([i] or [e]), all three singular exponents are considered equally likely. As a result, only the prior probabilities rather than the base likelihoods make a difference in the final probabilities.

$$\begin{aligned} p(i|u) &\propto p(u|i)p(i) = 0.\overline{3} * 0.\overline{3} = 0.\overline{1} \\ p(e|u) &\propto p(u|e)p(e) = 0.\overline{3} * 0.\overline{6} = 0.\overline{2} \\ \therefore p(i|u) &= 0.\overline{1}/(0.\overline{1} + 0.\overline{2}) = 0.\overline{3} \end{aligned} \tag{4.5}$$

In practice, of course, most useful models will fall somewhere between these two extremes of regularization. As I argue in the remainder of this chapter, experimental evidence from Icelandic and Polish speakers suggest that while prior probabilities play a major role in determining speakers' predictions, speakers are not doomed to recapitulate lexical frequencies: they can and do make limited but principled use of information in provided base forms.

This section has so far assumed that the prior probabilities at play are *empirical priors* which define a distribution over forms that mirrors their relative frequencies in the lexicon. At this point I note that this is not by any means the only way to define a prior distribution. One alternative worth discussing is the use of *uniform priors*, which ignore information about lexical frequencies. Given n candidates, a uniform prior distribution over them would assign each candidate a probability of $1.0/n$. In the case of this example, the $n = 2$ candidates are [i] and [e]. Like empirical priors, uniform priors are compatible with any degree of regularization of the likelihood term.[1] Equations 4.6 show the calculation of $p(i|u)$ under the assumption of maximal regularization and uniform priors. Note that only the frequencies of exponents themselves are relevant, and the fact that [e] is used in two classes as opposed to one does not affect the calculations.

$$\begin{aligned} p(i|u) &\propto p(u|i)p(i) = 0.3 * 0.5 = 0.15 \\ p(e|u) &\propto p(u|e)p(e) = 0.3 * 0.5 = 0.15 \\ \therefore p(i|u) &= 0.15/(0.15 + 0.15) = 0.5 \end{aligned} \tag{4.6}$$

To summarize this section, through choices of regularization parameters and types of priors, it is possible for a Bayesian model of inflectional morphology to describe various types of morphological grammars. When there is no regularization, the model predicts strict adherence to the implicational relationships, i.e. conditional probabilities, found in the lexicon, similar to a prescriptive or pedagogical view of inflectional morphology. Conversely, when regularization is maximized in the conditional likelihood term, this prevents the model from making any productive use of known base forms of the target lexeme, reducing inference to either an exercise in matching lexical frequencies (with an empirical prior) or assignment of equal probability mass to every possible derivative candidate (with a uniform prior). The following sections demonstrate that none of these extremes adequately explain the Icelandic or Polish experimental data, and that instead the models with the best explanatory power make use of empirical priors with a moderate amount of regularization.

[1] Without any regularization, $p(i|u)$ for this inflectional system would equal 1, as shown in 4.4, regardless of the prior type. However, this result depends crucially on one plural exponent having an unregularized conditional probability of 0. Otherwise, the prior will affect the final probabilities.

4.2 Assessing prior influence in Icelandic

The previous chapter described *wug* tests (Berko, 1958) on speakers of Icelandic and Polish, which were designed specifically to test the single surface base hypothesis and the base independence hypothesis, respectively. In both experiments, participants did not exhibit anywhere near perfect recapitulation of the strong—in some cases exceptionless—implicational relationships within their lexicons. This section revisits the results from the Icelandic experiment and develops the claim that participants' response patterns constitute more than just noise. Instead, I propose an explanation of participant behavior based on the concepts of Bayesian priors and regularization introduced earlier in this chapter. The explanatory power of this model of speaker behavior validates the Bayesian view of inflectional morphology with empirical priors, supporting the central equation of sublexical morphology.

This section assumes a familiarity with the experimental methodology described in 3.3.1. Discussion from this point onward assumes little about the nature of base reference in Icelandic, and so it is not necessary for readers to be familiar with the hypotheses that the experiment was designed to test. It should suffice instead to understand that the GenSg and NomPl forms of Icelandic nouns provide useful information—which speakers do indeed use—in predicting AccPl forms. Finally, note that this section presents strictly *post hoc* analysis of the experimental data.

4.2.1 Lexical frequencies in Icelandic

According to the hypothesis that Icelandic speakers make use of empirical priors in their inference of inflected forms, speakers' predicted probabilities of morphological exponents should measurably correspond to the proportions of those exponents in the Icelandic lexicon, especially in the absence of substantial information to the contrary. Determining the lexical frequencies of these exponents therefore constitutes the first step in assessing this hypothesis.

Type frequencies of AccPl exponents among Icelandic nouns were extracted from the *Database of Modern Icelandic Inflection* (Bjarnadóttir, 2012) using automated searches based on regular expressions, followed by manual checks performed by a native speaker to eliminate false positives. Proper nouns in Icelandic often inflect idiosyncratically, and so since Icelandic marks proper nouns with cap-

italization, words with capital letters were excluded. Moreover, because participants' responses in the experiment were limited to the four AccPl exponent choices *-ar*, *-a*, *-ir*, and *-i*, I limit the rest of this discussion to only forms with one of those endings. According to the regular expression search, the AccPl forms of fifty-five percent of noun lexemes in Icelandic end with one of these four sequences.[2]

The search procedure using regular expressions only produces counts of AccPl forms which end with particular sequences, regardless of whether or not these sequences constitute a suffix. Even under the assumption that Icelandic speakers make use of empirical priors based on lexical frequencies, there remains the question of whether these priors are based on frequencies of surface patterns, i.e. the presence of particular sequences of segments/characters in the AccPl, or based on frequencies of morphological exponents like suffixes themselves. If considering only true suffixes, then the procedure matching AccPl forms against regular expressions would overcount forms which take a null suffix but coincidentally end in the specified sequence.[3] Because there are no nouns in Icelandic with a null AccPl suffix but a non-null NomSg suffix, it is possible to exclude such cases by removing all lexemes whose AccPl form is identical to its NomSg form. I term the simple, surface-based regular expression search method the *surface* method, and the more complex method the *suffixal* method.

[2]To my knowledge there is no direct way to incorporate into statistical analyses the fact that response options were limited to exponents comprising only 55% of the lexicon. As this section shows, the frequencies within this 55% closely mirror participant responses, but perhaps some of the deviation from a perfect recapitulation of the lexical frequencies of *-ar*, *-a*, *-ir*, and *-i* is attributable to this issue.

[3]This procedure would also overcount forms ending in a suffix which coincidentally ends in the specified sequence, but there are no such nuisance suffixes in Icelandic for any of the four target suffixes.

	Surface	Suffixal
-a	31959 (43.2%)	31495 (52.3%)
-i	19892 (26.9%)	7260 (12.1%)
-ir	14637 (19.8%)	14415 (24.0%)
-ar	7445 (10.1%)	7060 (11.7%)

Figure 4.2: Counts of lexemes whose AccPl forms take each of the four target endings, based on the *Database of Modern Icelandic Inflection* (Bjarnadóttir, 2012). Data include *surface* counts, based on regular expression searches on AccPl forms, and *suffixal* counts, which include only AccPl forms bearing a non-null suffix as compared with its NomSg form.

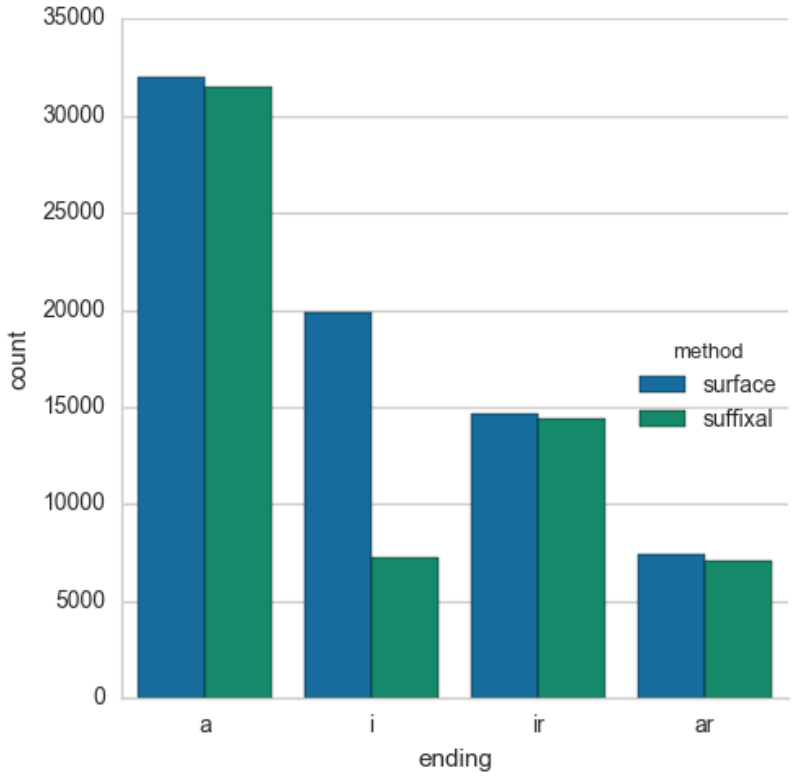

Figure 4.3: Visualized counts of lexemes whose AccPl forms take each of the four target endings, based on the *Database of Modern Icelandic Inflection* (Bjarnadóttir, 2012). Data include *surface* counts, based on regular expression searches on AccPl forms, and *suffixal* counts, which include only AccPl forms with a non-null suffix.

Figure 4.3 visualizes the results of these two counting procedures: surface counts of AccPl regular expression matches in blue on the left of each pair, and counts of true AccPl suffixes in green on the right of each pair. Counts for *-i* differ substantially between the two procedures—largely due to *-i*-final neuter AccPls with null suffixes—while counts for the other three suffixes differ minimally. As a result of these differences, not only the proportions but also the by-frequency orderings of the four suffixes vary between the two counting procedures. Specifi-

cally, whereas *-i* is the second-most frequent ending according to the raw counts, in the purely suffixal counts *-i* is barely more frequent than the least frequent ending, *-ar*.

4.2.2 Evidence for empirical priors in Icelandic

The most straightforward way to evaluate participants' prior distributions over the four available suffixes is to inspect how participants behaved when provided no novel information that might significantly influence their responses. The presentation condition in which lexemes were introduced using only their DatPl forms—that in which neither the GenSg nor NomPl was provided—meets this criterion. Note that while stem shape may also affect judgments about suffix appropriateness, the stimuli were designed and balanced so as to control for such effects; see 3.3 for further detail.

For comparison, we can inspect participants' response patterns in the presentation condition in which all base forms are provided, which in principle (see figure 3.8) should provide speakers the information necessary to virtually disqualify all but one suffix candidate. Because stimuli were balanced across the presentation conditions and the four inflectional classes, such "perfect" (*traditional*, in the sense introduced previously) behavior in the condition providing all base forms would predict that overall response counts should be even across the four possible suffixes.

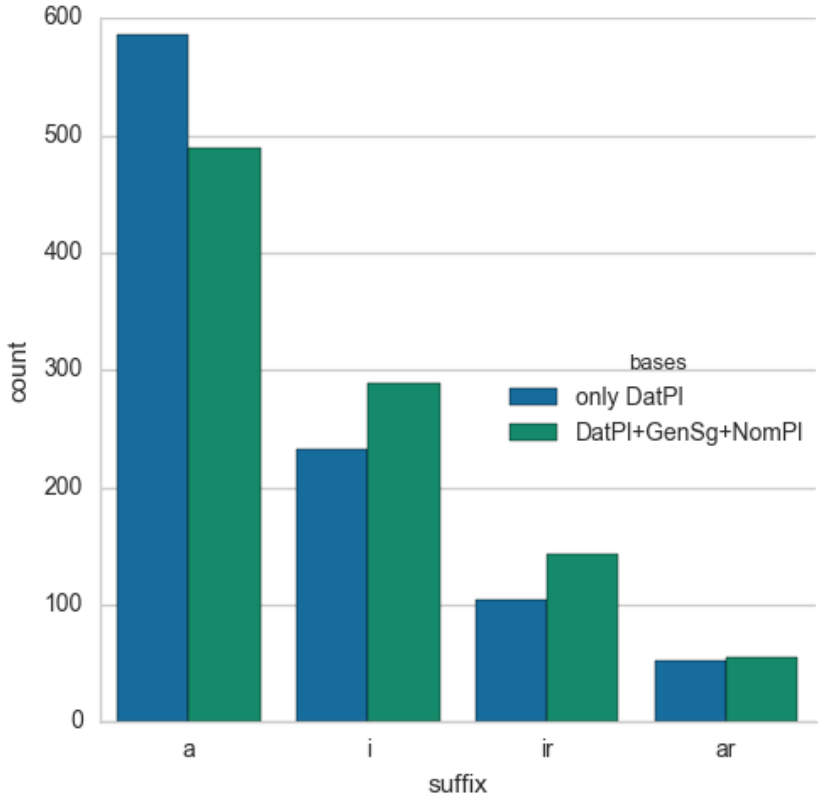

Figure 4.4: Frequencies of participants in the Icelandic experiment selecting an AccPl with each of the four possible suffixes. Blue bars on the left show response frequencies when only a lexeme's DatPl was provided, while the green bars on the right show responses when all three base forms were provided.

Figure 4.4 shows response patterns in both presentation conditions. Three conclusions can be drawn from these data. First, these responses are inconsistent with the hypothesis that Icelandic speakers make use of uniform priors. In the *only DatPl* presentation condition, which should reveal participants' prior beliefs about suffix distributions, rates of selecting each of the four suffixes differed substantially. Impressionistically speaking, rather than a uniform distribution, responses

constitute something closer to an exponential distribution, in which the frequency of *-a* responses is roughly twice as great as that of *-i* responses, which in turn is roughly twice as great as that of *-ir* responses, which finally is itself roughly twice as frequent as *-ar* responses. It would be implausible for a uniform prior distribution to generate the observed responses.

To validate this conclusion, and throughout the rest of this chapter, I use chi-squared tests for goodness of fit as implemented in the `chisq.test` function in R (R Core Team, 2013). In a general sense, this test is ideal for the task of determining the compatibility of a hypothetical distribution (such as a particular prior distribution over AccPl forms) with a set of responses ranging over the same values. This is because the test provides an estimate of the probability of the provided hypothetical distribution having generated the empirical data themselves. Chi-squared values for tests with d degrees of freedom are indicated as $\chi^2(d) = value$, and I use these values to arrive at p-values which indicate the likelihood of the data given a specified hypothetical distribution. I acknowledge, however, that one assumption of chi-squared tests is not met: responses are not all strictly independent from each other, since experiment participants each provided multiple responses and each lexeme was seen by multiple participants. Because there is to my knowledge no convenient test that is comparable to a chi-squared test while allowing for such groupings of data, I opt to report chi-squared values with this caveat, primarily as a quantitative, heuristic companion to qualitative evaluations. In the future, it may be useful (and better aligned with the arguments of this dissertation) to take an explicitly Bayesian approach here, using Monte Carlo markov chains to estimate data probabilities and performing Bayesian model comparison (Robert, 2007).

Returning to the specific question of whether speakers show evidence of employing uniform priors, I performed a chi-squared test of the experimental data for given probabilities [0.25, 0.25, 0.25, 0.25]. This test yields a chi-squared value of $\chi^2(3) = 709.7$ and a p-value of less than 0.001 for responses in the *only DatPl* condition. Therefore the response data in this presentation condition are highly unlikely to have been generated from a uniform underlying prior distribution.

Second, these response patterns are more compatible with the hypothesis that participants make use of empirical priors, specifically those based on frequencies of surface segment/character sequences, rather than on frequencies of true suf-

fixes. Qualitatively, the rate of *-i* responses in particular relative to other responses more closely resembles the distribution of *surface* counts than it does the distribution of *suffixal* counts. A chi-squared test for the lexical proportions of the four target sequences taking only true suffixes into account yields a chi-squared value of $\chi^2(3) = 227.42$, lower than that for the uniform distribution, but a p-value still less than 0.001. The same test on the proportions of AccPl forms ending in the target surface sequences (including null suffixes) yields a chi-squared value of $\chi^2(3) = 129.69$, the lowest of any distributions considered. However, the p-value for this test is still less than 0.001. From the chi-squared values, one can conclude that even though the p-values are all too small to directly compare, the *surface* distribution gives the highest likelihood to the experimental data.

The fact that even the best model's p-value is so low indicates that while the empirical prior based on frequencies of surface patterns rather than suffixes best corresponds to the experimental data, there are unaccounted-for sources of noise affecting participant responses. This may be due to the fact mentioned previously that the assumptions of chi-squared tests are not strictly met. Future research on this topic may benefit from taking a Bayesian approach to testing such questions.

As a sanity check, I also compared the fits of these three hypothetical distributions to the DatPl-only experimental data using Kullback-Leibler divergence (Kullback & Leibler, 1951). Kullback-Leibler divergence is an information theoretic measure that quantifies how well one probability distribution approximates another. Specifically, this measure indicates how much information (in the information theoretic sense) is lost by approximating one distribution using another. Thus this measure is useful as a second way of evaluating how well the various hypothetical prior distributions predict the experimental data. Since the theoretical foundations of this measure are quite different from those of the chi-squared test, I consider it a useful way to ensure that differences in chi-squared values are not attributable only to the the structure of that test itself.

As shown in Figure 4.5, the Kullback-Leibler divergences of the three hypothetical distributions from the observed distribution of responses also support the conclusion that the *surface* distribution best explains these responses.

Uniform	Empirical	
	Surface	Suffixal
0.341	0.069	0.118

Figure 4.5: Kullback-Leibler divergences of hypothetical prior distributions over AccPl endings from the observed response distribution from the Icelandic study in the presentation condition providing only DatPl forms.

Finally, based on the results in the *DatPl+GenSg+NomPl* presentation condition in which all bases were provided, it is clear that prior probabilities greatly influence speakers' judgments. A complete lack of regularization would predict evenly distributed responses in the condition in which all bases were provided. However, beyond the obvious qualitative mismatch between this prediction and the observed response patterns in the *DatPl+GenSg+NomPl* condition, a chi-squared test on these response frequencies and the uniform distribution [0.25, 0.25, 0.25, 0.25] yields a chi-squared value of 313.7 and a *p*-value of less than 0.001. However, the response patterns in this condition were more consistent with this uniform distribution than responses in the *only DatPl* condition, corroborating the finding from 3.3.2 that information in base forms shifts speaker judgments toward consistency with the conditional probabilities of AccPl forms in the lexicon. In the terms introduced earlier in this chapter, we can conclude that under a sublexical morphology interpretation of these results, Icelandic speakers' exhibit neither maximal regularization nor a complete lack of it, but rather some intermediate amount.

Among alternative causes of this response distribution that I have considered, none explain the experimental responses as well as the hypothesis that speakers' prior beliefs are quantitatively grounded in the surface frequencies of the target AccPl endings. One might consider the experimental context to unfairly promote an even distribution of responses among the four options; but while this may be the case to an extent, such an explanation cannot account for the varying response frequencies in the *only DatPl* condition. Given that of the four AccPl suffixes under consideration, *-a* is the most frequent in neologisms and loanwords (Gunnar

Ó. Hansson, p.c.), speakers should disproportionately prefer *-a* responses. While *-a* was the most common response in the experiment, this explanation cannot account for the non-negligible frequencies of other responses which, again, mirror lexical frequencies. However, although these alternative explanations cannot by themselves supplant the utility of empirical priors in modeling the experimental results, some combination of them may constitute part of the aforementioned noise which resulted in the chi-squared tests' low p-values.

4.3 Assessing prior influence in Polish

This section follows the methods of the previous one, revisiting the response data from the Polish experiment described in 3.4 and evaluating what information these data provide about Polish speakers' use of prior probabilities. The Polish patterns under investigation are in some ways simpler and in other ways more complex than the Icelandic patterns, and so the conclusions that can be drawn from the Polish experimental data differ somewhat from those that can be drawn from the Icelandic response data. Primarily, the findings of this section serve to support the findings in the previous section. Overall, the Polish data exhibit an influence of empirical priors very similar to that observed in the Icelandic data.

4.3.1 Lexical frequencies in Polish

In the Polish experiment, participants were asked to select their preferred NomPl form for each lexeme from just two choices: one form ending in *-e* and one ending in *-a*. Unlike in the Icelandic experiment, there is not a one-to-one correspondence between these endings and inflectional classes: the *-e* suffix is consistent with the soft masculine and soft feminine classes, while the *-a* suffix is consistent only with neuter classes. Figure 4.6 repeats the table from chapter 3. Note that while these NomPl suffixes may also correspond to non-soft inflectional classes, especially in the case of the neuter *-a*, the soft consonants ending each of the stems used as experimental stimuli should in principle force participants to consider the lexemes members of a soft declension; whether participants did so is one empirical question that this section addresses.

	GenSg	GenPl	NomPl
Soft neut.	-a	∅	-a
Soft masc.	-a	-y	-e
Soft fem.	-y	∅	-e

Figure 4.6: The suffixes associated with the GenSg, GenPl, and NomPl forms of soft neuter, masculine, and feminine nouns in Polish.

As in the case of Icelandic, there are multiple ways that lexical frequencies as usable by an empirical prior could be construed. The simplest way is to count the number of surface occurrences of *-e* and *-a* endings on Polish NomPl forms, and a somewhat more nuanced method excludes AccPl forms which actually have a null suffix but have a stem ending in one of these characters. I maintain the conventions of the previous section in calling these methods *surface* and *suffixal*, respectively. Additionally, in Polish there is a third count which may be relevant: that of *-e* and *-a* NomPl forms only among lexemes ending in a soft consonant, i.e. only within soft declensions. Since stem shapes make softness/hardness clear, such knowledge could plausibly be used in establishing participants' prior distributions over their response choices.

To determine these lexical frequencies, I extracted counts of common nouns from *PoliMorf* (Woliński *et al.*, 2012), the self-described "ultimate morphological resource for Polish" which builds off of the *Grammatical Dictionary of Polish* (*SGJP*) (Saloni *et al.*, 2007). Counts of *e*-final and *a*-final NomPl forms were produced by performing regular expression searches on the corpus. To arrive at counts excluding null suffixes, I compared NomPl forms to their lexemes' NomSg forms. As in Icelandic, there are no inflectional classes in Polish whose NomPl forms have null suffixes but whose NomSg forms have non-null suffixes, and so null-suffixed NomPl forms were excluded by removing NomPl forms which are identical to their lexemes' NomPl forms. Finally, counts of only *-e*-final and *-a*-final soft lexemes were extracted using regular expressions that combined the two target suffixes with the soft stem endings used in the experiment: *-ni*, *-si*, *-zi*, and *-ci*.[4] Figure 4.7 shows

[4]The soft consonants which end these stems are typically represented orthographically as *ń*, *ś*, *ź*,

the results of these searches.

	Surface	Suffixal	Soft
-e	28630 (77.4%)	22530 (74.0%)	4152 (61.3%)
-a	8346 (22.6%)	7908 (26.0%)	2626 (38.7%)

Figure 4.7: Counts of lexemes whose NomPl forms end in each of the target characters, based on *PoliMorf* (Woliński *et al.*, 2012). Data include *surface* counts based on regular expression searches on NomPl forms, *suffixal* counts which include only NomPl forms bearing a non-null suffix as compared with its NomSg form, and *soft* counts which include only NomPl forms whose stems end in one of the four soft consonants used in the Polish experiment.

and *ć*, respectively. However, according to the orthographical conventions of Polish, these sounds are written without a diacritic and with a following *i* when preceding a vowel. For example, the lexeme LOVE is spelled *miłość* in the NomSg and *miłości* in the NomPl.

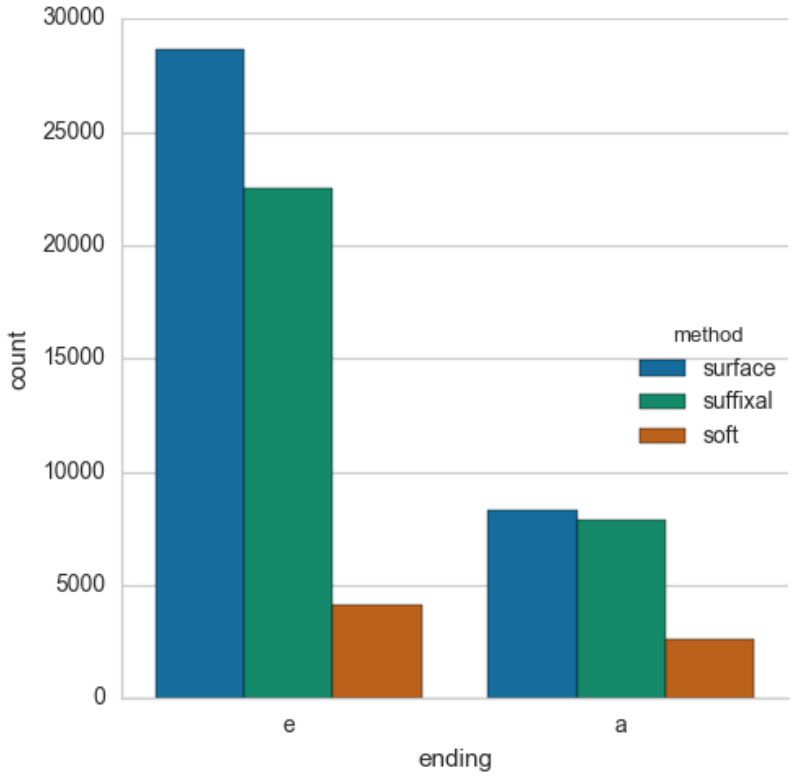

Figure 4.8: Visualized counts of lexemes whose NomPl forms end in each of the target characters, based on *PoliMorf* (Woliński *et al.*, 2012). Data include *surface* counts based on regular expression searches on NomPl forms, *suffixal* counts which include only NomPl forms bearing a non-null suffix as compared with its NomSg form, and *soft* counts which include only NomPl forms whose stems end in one of the four soft consonants used in the Polish experiment.

Figure 4.8 visualizes the results of these three counting procedures: surface counts of NomPl regular expression matches in blue on the left of each group, counts of true NomPl suffixes in green in the middle of each group, and in orange on the right side of each group, counts of only NomPl endings whose lexemes end in one of the four soft consonants used in the experiment. According to all

of these counting procedures, *-e* endings outnumber *-a* endings. This difference is most pronounced, however, in the surface counts of *-e*-final and *-a*-final NomPl forms, and least pronounced among the counts of soft lexemes. These results also demonstrate that only small minorities of the forms with these endings belong to a soft inflectional class, especially among *-e*-final forms.

4.3.2 Evidence for empirical priors in Polish

As in the preceding section, comparisons of Polish response patterns with lexical patterns bear on three central questions: whether Polish speakers make use of uniform or empirical priors, what types of lexical patterns empirical priors are based on (in the case that they are used at all), and how much regularization is at play in speakers' judgments. I proceed through these topics in the above order.

The most straightforward way to evaluate participants' prior distributions over the two available endings is, again, to inspect how participants behaved when provided no novel information that might significantly influence their responses. The presentation condition in which lexemes were introduced using only their DatPl forms—that in which neither the GenSg nor GenPl was provided—meets this criterion. Whether prior distributions are also affected by stem shape is discussed further below. For comparison, we can inspect participants' response patterns in the presentation condition in which all base forms are provided, which in principle (see figure 4.6) should provide speakers the information necessary to decide on a single candidate ending. Because stimuli were balanced across the presentation conditions and the three genders, such "perfect" behavior in the condition providing all base forms would predict response rates of approximately 67% for *-e* and 33% for *-a*. Figure 4.9 visualizes response rates in these two presentation conditions.

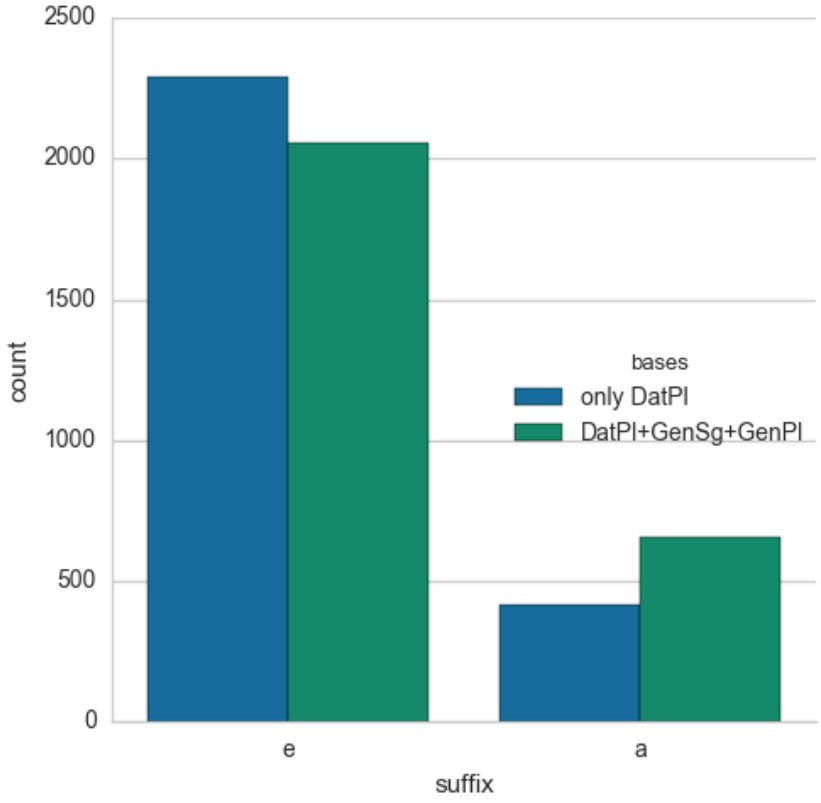

Figure 4.9: Frequencies of participants in the Polish experiment selecting a NomPl with each of the two possible suffixes. Blue bars on the left show response frequencies when only a lexeme's DatPl was provided, while the green bars on the right show responses when all three base forms were provided.

These response patterns are inconsistent with the hypothesis that speakers make no use of lexical frequencies when performing morphological inference, i.e. that they make use of uniform rather than empirical priors. A chi-squared test of the likelihood of the *only DatPl* experimental data given a uniform distribution assigning 0.5 probability to both *-e* and *-a* yields a chi-squared value of $\chi^2(1) = 1292.2$ and a p-value of less than 0.001. This indicates that it is highly unlikely that partic-

ipants with a uniform prior distribution over the two options would produce these response patterns.

Among the three possible interpretations of an empirical prior, the one making use only of surface frequencies of the endings *-e* and *-a* is most compatible with data from the *only DatPl* condition. While all three versions yielded p-values less than 0.001, the chi-squared values for the *surface*, *suffixal*, and *soft* distributions are $\chi^2(1) = 78.80$, $\chi^2(1) = 156.21$, and $\chi^2(1) = 616.38$, respectively. These results indicate that even though the p-values are too small to compare directly, the empirical prior model based on surface frequencies of the two endings gives the highest likelihood of the response patterns. As shown in Figure 4.10, the Kullback-Leibler divergence values of these hypothetical distributions from the response distribution further support these conclusions, indicating that the *surface* distribution is closest to the experimental response distribution. However, because the chi-squared test data likelihood even given the best model is so small, this interpretation suggests that there must be some unaccounted-for sources of noise in the results.

Uniform	Empirical		
	Surface	Suffixal	Soft
0.262	0.016	0.032	0.130

Figure 4.10: Kullback-Leibler divergences of hypothetical prior distributions over NomPl endings from the observed response distribution in the DatPl-only condition of the Polish study.

In addition, based on the results in the presentation condition in which all bases were provided, it is clear that prior probabilities greatly influence speakers' judgments. While overall response patterns in this condition do look qualitatively similar to the two-to-one odds predicted by the hypothesis that participants made perfect use of information in the provided base forms, participants' success rate in selecting the appropriate endings falls far short of the high success rate that this hypothesis predicts. Specifically, participants in this condition—which, based on implicational relationships in the lexicon, should allow perfect accuracy—selected the ending consistent with these relationships only 20.8% of the time (751 out of

3614 responses). A chi-squared test of response frequencies in this condition given the [$0.\overline{6}$, $0.\overline{3}$] distribution predicted if participants make perfect use of these implicational relationships, i.e. the response distribution predicted if there is no regularization at play in these judgments, yields a chi-squared value of $\chi^2(1) = 103.23$. However, performing the same test using the [0.774, 0.226] distribution of the best-performing empirical prior yields a much lower chi-squared value of just $\chi^2(1) = 3.64$. I conclude accordingly that since the prior distribution is a better fit to these data, there is some regularization bringing participants' responses closer to the empirical prior distribution of AccPl exponents, even though—as the tests in chapter 3 show—participants also make significant use of the relevant lexical co-occurrence patterns.

4.4 Summary and discussion

This chapter has served two purposes. First, it has offered a Bayesian interpretation of a traditional model of inflectional morphology, and has used this interpretation to set up a typology of inflectional productivity patterns predicted by various parameter values. Second, it has offered experimental evidence that bears on the question of how speakers make use of Bayesian priors in inflectional morphology.

Evidence from the Icelandic experiment introduced in the previous chapter supports three conclusions about the influence of prior distributions in inflectional inference, and evidence from the related Polish experiment corroborates these findings. According to this evidence, speakers make use of empirical priors when performing morphological inference. These priors correspond to the lexical frequencies of the surface-based phonological shapes corresponding to their derivative form options, rather than to frequencies of exponents like suffixes *per se* or frequencies of inflectional classes. Moreover, these priors exhibit a strong influence on morphological judgments even when they conflict with novel information provided by other inflected forms; in a model like sublexical morphology, this behavior indicates strong regularization on conditional likelihoods, but not enough so as to prevent any substantial use of information in provided base forms.

These results speak more generally to the notion of *productivity* in inflectional morphology. Between this chapter and the previous one, I have shown that even when speaker judgments diverge from those predicted by strong implicational rela-

tionships in the lexicon, such divergence is largely principled, deriving from simple lexical frequencies, and such divergence does not mean that speakers are making no productive use of those implicational relationships. For example, as Kawahara (2011, 2016) summarizes, in light of experimental findings that Japanese speakers often fail to "correctly" generalize Japanese verbal inflection to novel verbs (Batchelder, 1999; Griner, 2001; Vance, 1991, 1987), some linguists have concluded that the Japanese verbal system lacks productivity. The Bayesian view of productivity that I propose, in which real productivity can be partially hidden by the influence of prior probabilities, would not so hastily lead to this conclusion. I hope that this discussion encourages researchers to revisit questions of morphological productivity with an eye to how the concepts of priors and regularization can impact our assessment of whether patterns are indeed productive.

Lastly, I acknowledge that these *post hoc* investigations of evidence from my Icelandic and Polish experiments do not themselves constitute incontrovertible evidence for surface-based empirical priors. From the tests I have performed on these results, it appears that some unknown sources of noise are affecting speakers' judgments in addition to—or, perhaps, instead of—their empirical priors: the highest p-values I obtained from any of my chi-squared tests for goodness of fit were still lower than $2.2 * 10^{-16}$. Curiously, the more moderated response patterns in the all-bases conditions actually achieve better chi-squared values and KL-divergences than responses in the *only DatPl* conditions, suggesting that whatever other noise there may be, it has the anti-moderation effect of distributing responses less evenly. Additionally, one could argue that the Polish results in themselves prove little, since in Polish the expected response distribution given no regularization (the *traditional* model) is $[0.\overline{6},\ 0.\overline{3}]$, close to the observed response distribution. This is why I use the Polish results here primarily as a way of supporting the stronger claims made from the Icelandic study, in which the predicted response distributions vary substantially. In part because of these mitigating factors, I look forward to the possibility of future experiments designed specifically to assess the influence of empirical priors.

Chapter 5

Conclusion

This dissertation has presented and validated a novel approach to modeling the *paradigm cell filling problem*, that is, the task of inferring unknown forms within an inflectional paradigm. Humans perform this task in ways that evidence the learning of a complex generative morphological system, and yet investigations of this specific topic that combine formal modeling with experimental methods are rare. The goal of the research presented here has been to contribute to our collective understanding of how humans use their native languages' inflectional systems. By proposing not only a theoretical framework for understanding the use of inflectional morphology, but also a concrete implementation of the theory with a concomitant learning algorithm, I have laid the groundwork for future research in both theoretical linguistics and natural language processing.

This chapter concludes the dissertation with a summary of the previous chapters and further discussion. Section 5.1 reviews the general Bayesian view of inflectional morphology and my specific proposal of *sublexical morphology*, adding in-line references to the experiments I performed in order to validate the claims that make up my proposal. Section 5.2 then explores ways in which sublexical morphology may be of use to theoretical linguists beyond its core use case of "solving" the paradigm cell filling problem, including its potential for investigating paradigm leveling, paradigmatic gaps, and hypotheses about paradigm entropy. Finally, 5.3 addresses ways that follow-up research could address some limitations of the sublexical approach.

5.1 Summary of proposals and evidence

At the most fundamental level, I have proposed that we can conceive of the paradigm cell filling problem probabilistically, within a Bayesian framework whose variables are the surface-level inflected forms of lexemes, i.e. observed forms rather than abstract underlying representations. To "fill" a paradigm cell, a native speaker, knowing only a set of *base* forms of a lexeme, must infer and produce some theretofore unknown *derivative* form of the same lexeme. Under the probabilistic interpretation that I have proposed, filling a paradigm cell means not the selection of an optimal output form for the derivative, but rather the inference of a probability distribution over derivative forms followed by sampling from that distribution.

To formalize these ideas, I define a discrete distribution D which ranges over the possible forms of the derivative being inferred. (The set of possible forms is determined, for example, by the sublexicons of sublexical morphology.) This distribution is conditioned on a set of base form variables B, one for each base cell for which the speaker has observed a form of the lexeme, and each variable ranging over the observed forms of that lexeme in that base cell. Equation 5.1 shows the conditional probability distribution that a model of the paradigm cell filling problem generates.

$$p(D|B_1, B_2, ...B_n) \tag{5.1}$$

The first hypothesis about morphological inference that I placed under scrutiny was the *single surface base hypothesis* of Albright (2002) and subsequent papers. In brief, given the probabilistic interpretation that I proposed, this hypothesis constitutes a claim that inference about a derivative form makes use of only information contained in a single *privileged* base form. Equation 5.2 uses the notation developed so far to show the equality predicted by the single surface base hypothesis.

$$p(D|B_1, B_2, ...B_n) = p(D|B_{privileged}) \tag{5.2}$$

To assess this hypothesis, I carried out the experiment on Icelandic speakers described in the first part of chapter 3. Targeting a specific pattern within the Icelandic nominal inflection system, this experiment introduced a novel variation on

the *wug* test paradigm (Berko, 1958) to address the question of whether speakers are able to combine information from multiple bases in a single inference task. The experimental results are inconsistent with the single surface base hypothesis, suggesting that Icelandic speakers combine information from all available base forms when inferring unknown derivative forms. Despite the theoretical and computational appeal of the single surface base hypothesis, then, I concluded that speakers' grammars place no such limitation on their inferential capabilities.

Beyond simply proposing a probabilistic interpretation of the paradigm cell filling problem, perhaps the most essential claim I have made in this dissertation is that we can use Bayes's theorem to better understand and predict how speakers "solve" this "problem". As equation 5.3 shows, Bayes's theorem makes it possible to decompose the target conditional probability distribution into a *likelihood term* conditioned on the derivative and a *prior term*.

$$p(D|B_1, B_2, ...B_n) \propto p(B_1, B_2, ...B_n|D)p(D) \tag{5.3}$$

This manipulation itself does not lead to any particular increase in ease of modeling, because there are no clearly useful direct interpretations of the notions of a joint probability distribution over bases given a derivative $p(B_1, B_2, ...B_n|D)$ or the notion of a prior probability of a derivative $p(D)$; the probabilistic view of these distributions under-determines how they should be calculated. This is why I have gone one step further, introducing the framework of *sublexical morphology* which makes these distributions readily interpretable and calculable. According to sublexical morphology, the entire lexicon is partitioned into morphologically homogeneous sub-parts called *(paradigm) sublexicons*. Given a sublexicon and at least one base form, one can generate a derivative form for any derivative cell. Because of this near equivalence of sublexicons with derivative forms, sublexical morphology claims that distributions over derivative forms in the equations above can be replaced with distributions over sublexicons S, with derivative distributions then generable from sublexicon distributions. Essentially, morphological inference reduces to a straightforward classification problem in which selection of a sublexicon for some lexeme is tantamount to selecting its derivative form. Equation 5.4 summarizes these claims.

$$p(D|B_1, B_2, ...B_n) = p(S|B_1, B_2, ...B_n) \propto p(B_1, B_2, ...B_n|S)p(S) \qquad (5.4)$$

By substituting sublexicon distributions for derivative distributions, the distributions in this equation become both more intuitive and easier to calculate. As a demonstration of these properties, I will describe the sublexical interpretation of the two terms on the far-right-hand side of equation 5.4, first the likelihood term $p(B_1, B_2, ...B_n|S)$ and then the prior term $p(S)$.

The term $p(B_1, B_2, ...B_n|S)$ indicates the joint probability distribution over base forms given a sublexicon. Within sublexical morphology, the probability of a set of base forms given a sublexicon is determined by that sublexicon's Maximum Entropy harmonic grammar (Goldwater & Johnson, 2003; Hayes & Wilson, 2008), also called its *gatekeeper* grammar. These grammars are parameterized by weighted constraints, whose violation profiles over the provided base forms combine to result in an overall probability of the forms.

However, although the sublexical approach provides this method of calculating base likelihoods for each sublexicon, a system that models the entire joint probability distribution over base forms may pose a problem from the standpoint of learnability. Under the gatekeeper grammar interpretation of base likelihoods, for example, defining the joint space over base forms would require a massive proliferation of *cross-base constraint conjunctions*, constraints which refer to phonological material in multiple base forms at once. Since phonological constraint learning already poses substantial challenges even without this addition of orders of magnitude more complexity (Hayes & Wilson, 2008), it would be highly desirable if one could empirically determine that humans make use of only a small portion of this constraint space.

In order to evaluate whether such simplifications are empirically justified, I defined the *base independence hypothesis*, whereby the probabilities of base forms are conditionally independent of each other given a sublexicon. Equation 5.5 represents this hypothesis mathematically, using the definition of conditional independence. If the base independence hypothesis is valid, then determining the joint distribution over bases given a sublexicon becomes far easier; if calculated using a gatekeeper grammar, for example, the search space (and possible set of constraints

to evaluate during inference) would be restricted only to constraints which each evaluate some property of a single base form.

$$(B_1, B_2, ...B_n|S) = p(B_1|S)p(B_2|S)...p(B_n|S) \tag{5.5}$$

Seeking to test this hypothesis, I performed an experiment with native speakers of Polish, as described in the second part of chapter 3. This experiment used a methodology similar to that of the Icelandic experiment, but targeted a part of the Polish nominal system whose implicational relationships render it a viable testing ground for the base independence hypothesis. Participants' behavior in this experiment was consistent with the base independence hypothesis, although some unexplained irregularities in their overall response patterns make me wary of considering these results definitive. If the validity of the base independence hypothesis can be confirmed, e.g. by additional experimentation, then these findings will benefit theoretical and—especially—computational models of inflectional morphology, simplifying analyses for the former and facilitating learning and inference for the latter.

I turn now to the term in equation 5.4 indicating the prior probability distribution over sublexicons, $p(S)$. Chapter 4 contains the bulk of the discussion of this aspect of the model. In sublexical morphology, this distribution forms an *empirical* prior distribution matching the relative "sizes" of the various sublexicons in an inflectional system. The precise manner in which the size of a sublexicon is quantified, however, was not known *a priori*, nor was there any particular theoretical reason to define it in some particular way. Moreover, there was no empirical evidence that prior distributions in sublexical morphology should be based on lexical frequencies at all.

Intending to assess whether speakers do indeed make use of empirical priors, and to determine how speakers arrive at them, I revisited the results of the Icelandic and Polish experiments introduced originally in chapter 3. These analyses of participant responses were performed on a strictly *post hoc* basis, but taken together they suggest that speakers of Icelandic and Polish make use of empirical priors when performing morphological inference, and moreover that the influence of these priors largely—but not completely—overshadows the effect of base forms

on their posterior distributions over derivative candidates. More specifically, these analyses suggest that speakers' empirical priors reflect the surface frequencies of segment/character sequences associated with each derivative candidate, rather than the frequencies of exponents (e.g. suffixes) themselves or frequencies of inflectional classes.

The theoretical and experimental findings of this dissertation validate my proposal of a probabilistic, Bayesian approach to inference in inflectional morphology (the paradigm cell filling problem), demonstrating in particular the explanatory and predictive power of sublexical morphology. This research also constitutes the foundation of further research on probabilistic models of inflectional morphology and, more generally, on the formal limits on human abilities to perform such inference.

5.2 Other applications of sublexical morphology

Sublexical morphology directly models the paradigm cell filling problem, but its usefulness within linguistic theory extends beyond this scope. This section surveys three specific research topics within theoretical morphology about which sublexical morphology may help yield new and valuable insights. These topics include diachronic phenomena—paradigm leveling and the emergence of paradigmatic gaps—and hypotheses about the nature of predictability in inflectional systems—paradigmatic gaps and paradigm entropy conjectures.

5.2.1 Paradigm leveling

As discussed in chapter 3, Albright (2002), Albright (2008), and Albright (2010) among others have argued that historical patterns of *paradigm leveling* in Latin and Yiddish can be explained and predicted by the *single surface base hypothesis*. This hypothesis states that for any inflectional system, speakers can use only the form in a single *privileged* base cell to generate unfamiliar derivative forms, and that this privileged base cell is the cell whose forms are most informative about the forms in other cells, given the implicational relationships in the lexicon. The single surface base hypothesis accurately predicts the directionality of paradigm leveling in several cases discussed by Albright: the limitation to use of only a single base form means that phonological distinctions among inflectional classes which are only present in non-privileged cells are the ones at risk of diachronic loss. However, the results of the Icelandic experiment described in this dissertation

refute the claim that this limitation to a single base form holds cross-linguistically.

While it may appear then that I have traded an explanation of one phenomenon (paradigm leveling) for an explanation of others (those discussed in this dissertation) by falsifying the single surface base hypothesis, I propose that sublexical morphology may in fact also offer an explanation of at least some observed paradigm leveling patterns. In this subsection, I review the definition of paradigm leveling with reference to a standard instance of the phenomenon, the Latin HONOR "analogy", and then I demonstrate that even with only the mechanisms introduced in previous chapters, sublexical morphology is able to predict the directionality of this historical change.

	Old Latin		Golden Age Latin	
	Class 1 (high freq.)	Class 2 (low freq.)	Class 1 (high freq.)	Class 2 (low freq.)
NomSg	soror	hono***s***	soror	hono***r***
GenSg	sororis	honoris	sororis	honoris

Figure 5.1: A schematic of Old Latin and Golden Age Latin NomSg and GenSg forms relevant to the leveling of HONOR-like words. The forms for SISTER and HONOR are used as examples of forms in classes 1 and 2, respectively.

Figure 5.1 illustrates the key facts in Old Latin, which preceded this case of paradigm leveling, and in Golden Age Latin, which was spoken after the leveling (Albright, 2002). While there was some individual variation among lexical items, in general there are two classes of nouns thought to be relevant to the phenomenon: a group which I label as *Class 1*, which included many lexical items in Old Latin including the SISTER word *soror, sororis*, and a group which I label as *Class 2*, which included fewer lexical items including the HONOR word *honos, honoris*. Here I use the forms of those two exemplar lexical items to show the patterns of those classes in general. Note that while I have listed GenSg forms, the relevant contrast is more properly NomSg versus the *oblique* forms, which include GenSg and all other non-NomSg forms.

Crucially, the [-s] suffix that was the exponent of the NomSg in Class 2 in Old Latin became an [-r] suffix in Golden Age Latin, rendering the two classes morphologically equivalent in the more recent language. Because this change occurred among words like HONOR, and because these NomSg forms appear to have been rebuilt analogously to the NomSg forms of Class 1, this phenomenon is called the Latin HONOR analogy. Moreover, because this historical change resulted in a neutralization of a prior contrast between inflectional classes, it is an example of paradigm leveling.

There are two key questions about the Latin facts which any theory of paradigm leveling must address. First, why was it the NomSg form that changed instead of the GenSg form? It is conceivable that instead of the Class 2 NomSg forms changing, the Class 2 GenSg forms could have ended up taking a [-sis] suffix. Second, why did Class 2 forms change instead of Class 1 forms? It is also conceivable that SISTER-like words could have been leveled, taking the [-s] suffix of Class 2 words, rather than the other way around as was observed.

The sublexical morphology view of how paradigm leveling might emerge offers explanations for both of these directionalities of change. For an inflectional class to be leveled, i.e. undergo neutralization with some other class, its inflected forms which distinguish it from the class to which it levels must at some point be produced with the morphology of the leveled-to class. In other words, speakers faced with the paradigm cell filling problem and needing to infer these forms infer "incorrectly" that these forms take the morphology of a class other than the one to which they originally belonged. Once this process begins, these novel inferred forms are presumably heard by other speakers, are memorized by them, and then propagate with a decreasing need for the (mis-)inferential process. In the Latin case, for example, perhaps some speaker(s) innovated the form [honor] and similar forms for other words in its class, and these novel forms spread throughout the community of speakers.

In response to the first question, sublexical morphology predicts accurately that NomSg forms rather than GenSg forms would be altered by paradigm leveling. Using morphological operations plausibly learned by the algorithm described in section 2.5, the Old Latin sub-paradigm in Figure 5.1 would need only one operation to derive GenSg forms from NomSg forms, but two operations (one for each class)

to derive NomSg forms from GenSg forms. Figure 5.2 shows these operations.

	Sublexicon 1 (high freq.)	Sublexicon 2 (low freq.)
NomSg → GenSg	final segment → [ris]	final segment → [ris]
GenSg → NomSg	final [ris] → [***r***]	final [ris] → [***s***]

Figure 5.2: The morphological operations deriving NomSg and GenSg forms from each other in the paradigm sublexicons of Old Latin.

Under the assumptions of sublexical morphology, the selection of a sublexicon stands as a proxy for the generation of a derivative. For the "wrong" derivative to be produced, all that is required is for the lexeme to be associated with the "wrong" sublexicon. As Figure 5.2 makes clear, when deriving a GenSg from a NomSg, it does not matter at all which sublexicon a speaker considers a lexeme like HONOR to belong to; both sublexicons would result in the same derivative form, one ending in [-ris]. When generating a NomSg form from a GenSg form, however, the choice of sublexicon matters a great deal: this choice is equivalent to the choice between an [-r] suffix and an [-s] suffix on the NomSg. Paradigm leveling amounts to the consolidation of two sublexicons into one, and with the sublexicons shown in the figure above, such consolidation would only produce noticeable changes in inflected forms among NomSg forms, not among GenSg forms.

As for the question of why the class 2 forms changed rather than the class 1 forms, a sublexical morphology account of paradigm leveling would attribute this fact to the prior probabilities of the two sublexicons. It is no coincidence, in this account, that the forms in the less frequent class 2 were rebuilt to be more similar to the forms in the more frequent class 1. If a speaker of Old Latin relied on empirical priors as defined according to the empirical results from chapter 4, and especially if speakers relied on these priors to the extent that participants in the Icelandic and Polish studies appear to have, then we would expect the morphological operation of class/sublexicon 1 to be (mis-)applied to lexemes in class 2 far more often than the morphological operation of class/sublexicon 2 being (mis-)applied to lexemes in class 1. This behavior would result in speakers producing forms like

[honor] frequently and [soros] much less so, feeding the canonicalization which, by conjecture, results diachronically in paradigm leveling.

Taken as a whole, sublexical morphology successfully predicts both aspects of the directionality of the Latin HONOR analogy, at least as viewed as narrowly concerning the two noun classes mentioned here. A fuller account of the analogy based on sublexical morphology would need to demonstrate that the theory does not predict other NomSg–GenSg ambiguities in Old Latin yielding leveling changes—although the probabilistic nature of sublexical morphology's predictions make it difficult to clearly falsify using historical data. In general, my proposals here predict that paradigm leveling should be more likely in cases where two sublexicons (or inflectional classes) exhibit both a severe difference in lexical frequencies and a lack of robust phonological differences. Additionally, while I do not intend to suggest that sublexical morphology alone can account for all cases of paradigm leveling, it goes without saying that in order to constitute a universal theory of paradigm leveling, the framework would need to be tested on other datasets, especially the Yiddish dataset on which Albright's (2002 et seq.) model performs so uncannily well. I hope that this brief discussion of Latin can serve as the seed of future research into Bayesian and sublexical interpretations of paradigm leveling.

5.2.2 Paradigmatic gaps

Different strands of research on paradigmatic gaps have converged on the finding that such gaps correspond to parts of an inflectional system exhibiting a lack of predictability (Albright, 2003, 2009; Hansson, 1999; Sims, 2006). By paradigmatic gaps, I refer to logically possible forms of lexemes in an inflectional system which speakers avoid producing; for example, some verbs in Spanish including [asir] GRASP have no (canonical) first person singular present indicative form. Research on paradigmatic gaps typically addresses the question of why such gaps come about in the first place, as well as the question of why gaps occur in some parts of a paradigm but not others.

Directly or indirectly, the papers on paradigmatic gaps that I have cited here focus in part on the hypothesis that paradigmatic gaps tend to occur in less predictable parts of an inflectional system. The predictability of a derivative cell is defined in terms of how useful the morpho-phonological sub-regularities in other

cells of the paradigm would be when used to infer forms in that derivative cell. Perhaps, for example, even with knowledge of other forms of the lexeme GRASP in Spanish, speakers are unable to confidently predict a single most viable candidate for its first person singular present indicative form. Probabilistically, one could describe this situation by defining a probability distribution over the candidates for this derivative form, conditioned on the known base forms of the lexeme. If there is no single obvious "winning" candidate, then probability mass is distributed more evenly among the candidates in this distribution. In the terminology of information theory, this distribution is highly *entropic*.

Under a Bayesian view of inflectional morphology, the paradigm cell filling problem amounts to the inference of exactly such a conditional probability distribution. Because an entropy value can be calculated from any probability distribution, sublexical morphology can therefore be used to calculate the conditional entropy of any cell for any lexeme in an inflectional system. This property of sublexical morphology—especially using the implementation and learning algorithm described in chapter 2—makes it possible to directly test hypotheses about paradigmatic gaps and the entropy of individual cells in a paradigm. For example, a researcher could continue in the footsteps of Sims (2006) by performing a multi-base wug test like those described in this dissertation, but with a "decline to respond" answer option, and then test whether participants rely more on this extra answer option when the entropy of an inference task is high. Unlike other methods for calculating the predictability of parts of inflectional paradigms, sublexical morphology takes into account the prior probabilities of derivative forms, potentially making its estimates of entropy more reliable. A follow-up study could also test whether or not prior probabilities play a role in experimental participants' likelihood to decline to respond.

5.2.3 Paradigm entropy

Moving beyond the properties of individual cells in a paradigm, Ackerman & Malouf (2013) have performed information theoretic analyses of inflectional paradigms across a variety of languages, showing that while some information theoretic measures vary widely from language to language, others cluster tightly cross-linguistically, and that these results have interesting and useful theoretical consequences. For ex-

ample, they define a paradigm's *average conditional entropy* as "the average uncertainty in guessing the realization of one randomly selected cell in the paradigm of a lexeme given the realization of one other randomly selected cell."

Just as for paradigmatic gaps, sublexical morphology could be useful to researchers interested in testing these and other variations of hypotheses about the information theoretic properties of entire inflectional paradigms. Since the measures discussed by Ackerman & Malouf (2013) can all be derived from the probability distributions for lexemes' individual cells conditioned on some subset of those lexemes' other forms, a model of the paradigm cell filling problem like sublexical morphology can indirectly calculate these measures. Moreover, the Bayesian character of sublexical morphology also allows researchers to evaluate the impact of prior probabilities on these entropy values.

5.3 Limitations and future directions

The simplicity and utility of sublexical morphology derive mainly from its establishment of a deterministic mapping from sublexicon to derivative form via the morphological operations of each sublexicon. Because of this property, assigning a probability to a sublexicon is equivalent to assigning a probability to its corresponding derivative candidate. However, the price of this system is that a sublexicon must be completely homogeneous in terms of its morphological operations. This strict homogeneity requirement means that lexemes which differ morphologically in even a single respect (i.e. a single cell) must belong to separate sublexicons.

There are two inter-related negative consequences of this property of sublexical morphology models. The first is that sublexical morphology's intuitively overzealous partitioning of the lexicon could make it difficult for gatekeeper grammars to accurately assess the characteristic phonological properties of each sublexicon, weakening the gatekeepers' empirical accuracy. For example, in an extreme case, there might be two large classes of nouns which take the same morphological exponents (same within class, different between classes) *except* in one particular cell, in which each lexeme in the two classes has its own idiosyncratic exponent. In such a situation, there would need to be a sublexicon of size one (i.e. with only one associated lexeme) for every lexeme. Because a gatekeeper grammar's constraint weights are calculated by comparing the forms in its sublexicon to all other forms, this pro-

liferation of sublexicons would in all likelihood prevent the general phonological properties distinguishing the two greater classes from being effectively encoded in the constraint weights.

This problem mirrors a common tension in descriptive morphology: whether two lexemes which are mostly morphologically homogeneous ought to be considered members of the same or separate inflectional classes. Ambiguous cases abound, including Latin 3rd conjugation verbs vs. "3rd -*io*" verbs, and how Spanish verbs exhibiting diphthongization should be distinguished from those without. I take the long-standing difficulty of this problem as evidence that there may be no simple solution (although see e.g. Brown & Hippisley 2012 for work in this direction). Even so, one could conceive of a version of sublexical morphology in which there is a "soft" requirement of morphological homogeneity in a sublexicon, so that the grammar can include fewer sublexicons (and therefore, perhaps, more useful constraint weights) at the cost of guessing randomly—according, for example, to lexical frequencies—when needing to select an exponent for the heterogeneous cells merged into a single sublexicon. A more complex but potentially more powerful solution might treat sublexicon membership hierarchically, so that lexemes that are morphologically identical in some cells are treated as belonging to the same sublexicon at some level of a hierarchy, but are then split into separate sublexicons at a lower level due to differences in other cells. In such a system, weighted constraint violations could be summed along each path down the hierarchy to evaluate sublexicon probabilities.

Sublexical morphology's lack of phonology-driven processes to derive different surface realizations of forms in the same sublexicon exacerbates this problem. For example, sublexical morphology has no mechanism for using English phonotactics to derive the regular [-s], [-z], and [-ɪz] plurals from the same sublexicon. Although the appeal of sublexical morphology stems in large part from its lack of a need to learn global phonotactics and abstract underlying representations, it is possible that some limited abstraction of stored forms could help unify sublexicons. This approach resembles the "bundling" process described by Moore-Cantwell & Staubs (2014), which could be generalized from modeling pairs of cells to modeling entire inflectional systems much as sublexical morphology has generalized sublexical phonology (Allen & Becker, in review; Gouskova & Newlin-Łukowicz,

2013).

The second, related consequence of the homogeneous sublexicons requirement is that the problem of sublexicon proliferation described above becomes more serious as the number of cells being modeled increases. As the number of cells increases, the odds that two lexemes will be (perhaps undesirably) split into separate sublexicons increases commensurately. This is one reason that I have limited my discussions so far to relatively small inflectional systems and parts of inflectional systems. The solutions described above may help alleviate this problem as well.

However, truly massive inflectional systems, including "agglutinative" inflectional systems, pose further problems for sublexical morphology. From the standpoint of learnability, because every pair of cells' base sublexicon divisions must be learned before paradigm sublexicons can be inferred, as the number of cells n increases, the time it takes to learn the inflectional system's paradigm sublexicons increases at a rate of at least n^2. Additionally, sublexical morphology treats each cell as equally distinct from each other cell, and it therefore has no way of recognizing similarities among cells that may be useful. For example, the operation that maps Japanese past tense forms onto past conditional forms (concatenation of [-ra]) is the same regardless of whether a verb's polarity is affirmative or negative, or whether or not it is a conditional verb, etc.

I can think of no way of surmounting these issues within the general assumptions of sublexical morphology except by learning groupings of cells. For example, the learning algorithm could discover the identity relationship described in the previous paragraph, and could thereafter treat all past tense cells as a single "super-cell" for purposes of deriving past conditional forms. Such consolidation could proceed, for example, by finding pairs of cells with only one base sublexicon between them bidirectionally. However, this approach would still require that base sublexicons be learned for every pair of cells. As a potential workaround, the grammar might initially treat all cells as belonging to the same "super-cell" and only split off cells when required by data encountered by the learner.

Finally, the procedures for learning sublexicons that I have described in this dissertation have generally assumed that all inflected forms (within the part of an inflectional system being learned) are present in the training data for all lexemes. The need for this unrealistic assumption stems from the fact that when only a subset

of inflected forms are available, it may be unclear which paradigm sublexicon(s) a lexeme in the training data belongs to. A more sophisticated learning algorithm would need to deal with this ambiguity, perhaps by allowing lexemes to be associated probabilistically with as many sublexicons as appropriate. Calculation of derivative probabilities would therefore require marginalization across sublexicon membership probabilities.

Bibliography

ACKERMAN, FARRELL, JAMES P. BLEVINS, & ROBERT MALOUF. 2009. Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. In *Analogy in grammar: Form and acquisition*, 54–82. Oxford: Oxford University Press. → pages 3

——, & ROBERT MALOUF. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 89.429–464. → pages 139, 140

ALARCOS LLORACH, EMILIO. 1994. *Gramática de la lengua española*, volume 61. Madrid: Espasa Calpe. → pages 2, 13

ALBRIGHT, ADAM, 2002. *The identification of bases in morphological paradigms*. University of California, Los Angeles dissertation. → pages 21, 52, 53, 55, 57, 58, 103, 130, 134, 135, 138

——. 2003. A quantitative study of Spanish paradigm gaps. In *West Coast Conference on Formal Linguistics 22 Proceedings*, ed. by G. Garding & M. Tsujimura, 1–14, Somerville. Cascadilla. → pages 138

——. 2008. Explaining universal tendencies and language particulars in analogical change. In *Language universals and language change*, ed. by Jeff Good, 144–181. Oxford: Oxford University Press. → pages 53, 57, 58, 134

——. 2009. Lexical and morphological conditioning of paradigm gaps. In *Modeling Ungrammaticality in Optimality Theory*, ed. by Curt Rice & Sylvia Blaho. London: Equinox. → pages 138

——. 2010. Base-driven leveling in Yiddish verb paradigms. *Natural Language & Linguistic Theory* 28.475–537. → pages 134

——, & BRUCE HAYES. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 workshop on*

morphological and phonological learning, volume 6, 58–69. Association for Computational Linguistics. → pages 9, 49, 58

——, & BRUCE HAYES. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90.119–161. → pages 58

——, & BRUCE HAYES. 2011. Learning and learnability in phonology. In *The handbook of phonological theory*, ed. by John Goldsmith, Jason Riggle, & Alan Yu, 661–690. Hoboken: Wiley-Blackwell. → pages 10

ALLEN, BLAKE, & MICHAEL BECKER, in review. Learning alternations from surface forms with sublexical phonology. → pages 9, 17, 24, 30, 41, 42, 43, 49, 141

ANDERSON, STEPHEN R. 1992. *A-morphous morphology*. Cambridge: Cambridge University Press. → pages 5, 47

ARCHANGELI, DIANA, & DOUG PULLEYBLANK. 2012. Emergent phonology: evidence from English. In *Issues in English linguistics*, ed. by Ik-Hwan Lee, Young-Se Kang, Kyoung-Ae Kim, Kee-Ho Kim, Il-Kon Kim, Seong-Ha Rhee, Jin-Hyung Kim, Hyo-Young Kim, Ki-Jeong Lee, Kye-Kyung Kang, & Sung-Ho Ahn, 1–26. Seoul: Hankookmunhwasa. → pages 10

ÁRNASON, MÖRÐ UR. 2007. *Íslensk orðabók. 4th edition*. Reykjavík: . → pages 63

ARONOFF, MARK. 1994. *Morphology by itself: Stems and inflectional classes*. Number 22 in Linguistic Inquiry Monographs. Cambridge: MIT press. → pages 5, 47

BARR, DALE J, ROGER LEVY, CHRISTOPH SCHEEPERS, & HARRY J TILY. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68.255–278. → pages 74

BATCHELDER, ELEANOR OLDS. 1999. Rule or rote? Native-speaker knowledge of Japanese verb inflection. In *Proceedings of the Second International Conference on Cognitive Science*, 141–146. → pages 7, 128

BATES, DOUGLAS, REINHOLD KLIEGL, SHRAVAN VASISHTH, & HARALD BAAYEN. 2015a. Parsimonious mixed models. *arXiv preprint arXiv:1506.04967* . → pages 74

——, MARTIN MÄCHLER, BEN BOLKER, & STEVE WALKER. 2015b. Fitting linear mixed-effects models using lme4. *Journal of statistical software* 67.1–48. → pages 73, 96

BEARD, ROBERT. 1995. *Lexeme-morpheme base morphology: a general theory of inflection and word formation*. Albany: SUNY Press. → pages 5, 47

BECKER, M., A. NEVINS, & J. LEVINE. 2012. Asymmetries in generalizing alternations to and from initial syllables. *Language* 88.231–268. → pages 7

BECKER, MICHAEL, & MARIA GOUSKOVA, 2013. Source-oriented generalizations as grammar inference in Russian vowel deletion. Ms. lingbuzz/001622. → pages 7, 17

——, & JONATHAN LEVINE, 2012. *Experigen - an online experiment platform*. Available at https://github.com/tlozoot/experigen. → pages 67, 89

BERKO, J., 1958. *The child's learning of English morphology*. Radcliffe College dissertation. → pages 8, 51, 111, 131

BICKEL, PETER J, BO LI, ALEXANDRE B TSYBAKOV, SARA A VAN DE GEER, BIN YU, TEÓFILO VALDÉS, CARLOS RIVERO, JIANQING FAN, & AAD VAN DER VAART. 2006. Regularization in statistics. *Test* 15.271–344. → pages 108

BISHOP, CHRISTOPHER M. 2006. *Pattern recognition and machine learning*. New York: Springer. → pages 46

BJARNADÓTTIR, KRISTÍN. 2012. The Database of Modern Icelandic Inflection (Beygingarlýsing íslensks nútímamáls). In *Proceedings of Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL 8 / AfLaT 2012)*, ed. by Guy De Pauw, Gilles-Maurice de Schryver, Mikel L. Forcada, Kepa Sarasola, Francis M. Tyers, & Peter W. Wagacha, 13–18, Istanbul. European Language Resources Association (ELRA). → pages 63, 64, 65, 69, 111, 113, 114

BLEVINS, JAMES. 2006. Word-based morphology. *Journal of Linguistics* 42.531–573. → pages 47

BONAMI, OLIVIER, & GILLES BOYÉ. 2007. French pronominal clitics and the design of Paradigm Function Morphology. In *Proceedings of the Fifth Mediterranean Morphology Meeting*, 291–322, Bologna. → pages 47

BROWN, DUNSTAN, & ANDREW HIPPISLEY. 2012. *Network morphology: A defaults-based theory of word structure*. Cambridge: Cambridge University Press. → pages 9, 37, 49, 141

CHILDS, G. TUCKER. 2003. *An introduction to African languages*. Amsterdam: John Benjamins Publishing. → pages 24

CHOMSKY, N., & M. HALLE. 1968. *The sound pattern of English*. New York: Harper & Row. → pages 4

CHOMSKY, NOAM. 1956. Three models for the description of language. *IRE transactions on information theory* 2.113–124. → pages 4

——. 1957. *Syntactic structures*. The Hague/Paris: Mouton. → pages 4

——. 1995. *The minimalist program*. Cambridge: MIT Press. → pages 4

COLEMAN, JOHN, & JANET PIERREHUMBERT. 1997. Stochastic phonological grammars and acceptability. *arXiv preprint cmp-lg/9707017* . → pages 4

DALAND, ROBERT. 2015. Long words in maximum entropy phonotactic grammars. *Phonology* 32.353–383. → pages 25

DREYER, MARKUS, & JASON EISNER. 2011. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 616–627, Edinburgh. Supplementary material (9 pages) also available. → pages 50

EDDINGTON, DAVID, REBECCA TREIMAN, & DIRK ELZINGA. 2013. The syllabification of American English: Evidence from a large-scale experiment part I. *Journal of quantitative linguistics* 20.75–93. → pages 7

EINARSSON, STEFÁN. 1949. *Icelandic: grammar, text and glossary*. Baltimore/London: The Johns Hopkins University Press. → pages 61

FRIEDMAN, JEROME, TREVOR HASTIE, SAHARON ROSSET, ROBERT TIBSHIRANI, & JI ZHU. 2004. Discussion of boosting papers. *Annals of statistics* 32.102–107. → pages 108

GALES, MARK JF, KATE M KNILL, ANTON RAGNI, & SHAKTI P RATH, 2014. Speech recognition and keyword spotting for low resource languages: Babel project research at CUED. → pages 5

GOLDWATER, SHARON, & MARK JOHNSON. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory*, ed. by J. Spenader, A. Eriksson, & Ö. Dahl, 111–120, Stockholm. Department of Linguistics. → pages 10, 24, 30, 45, 132

GOUSKOVA, MARIA, & LUIZA NEWLIN-ŁUKOWICZ, 2013. Phonological selectional restrictions as sublexical phonotactics. Manuscript. → pages 9, 17, 41, 49, 141

GRINER, BARRY DAVID, 2001. *Productivity of Japanese Verb Tense Inflection: A Case Study*. University of California, Los Angeles dissertation. → pages 128

HANSSON, GUNNAR ÓLAFUR. 1999. 'When in doubt...': intraparadigmatic dependencies and gaps in Icelandic. In *Proceedings of the 30th Meeting of the North East Linguistic Society*, volume 29, 105–120. → pages 138

HANSSON, GUNNAR ÓLAFUR. 2006. Málfræðirannsóknir á öld upplýsingatækninnar – lítil reynslusaga. *Lesið í hljóði fyrir Kristján Árnason sextugan 26.desember 2006* . → pages 69

HAYES, B., & Z.C. LONDE. 2006. Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology* 23.59–104. → pages 7

——, & C. WILSON. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.379–440. → pages 4, 10, 16, 24, 30, 45, 86, 132

HAYES, BRUCE. 2011. Interpreting sonority-projection experiments: the role of phonotactic modeling. In *Proceedings of the 17th International Congress of Phonetic Sciences*, 835–838, Hong Kong. City University of Hong Kong. → pages 108

——. To appear. Comparative phonotactics. *Proceedings of the 50th Meeting of the Chicago Linguistic Society* . → pages 28

——, PÉTER SIPTÁR, KIE ZURAW, & ZSUZSA LONDE. 2009. Natural and unnatural constraints in hungarian vowel harmony. *Language* 85.822–863. → pages 7

——, & JAMES WHITE. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44.45–75. → pages 10

HLAVAC, MAREK. 2013. stargazer: LaTeX code and ASCII text for well-formatted regression and summary statistics tables. *URL: http://CRAN.R-project.org/package=stargazer* . → pages 76, 98, 100

HONRUBIA, J.L.C., J.L. CIFUENTES, & S.R. ROSIQUE. 2011. *Spanish Word Formation and Lexical Creation*. IVITRA research in linguistics and literature. Amsterdam: John Benjamins Publishing Company. → pages 13

HUDSON KAM, CARLA, & ELISSA NEWPORT. 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language learning and development* 1.151–195. → pages 4, 10

JAROSZ, GAJA. 2005. Polish yers and the finer structure of output-output correspondence. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, volume 31, 181–192, Berkeley. University of California Press. → pages 90

JESNEY, KAREN, & ANNE-MICHELLE TESSIER. 2009. Gradual learning and faithfulness: consequences of ranked vs. weighted constraints. In *Proceedings of the North East Linguistic Society 38*, ed. by Anisa Schardl, Martin Walkow, & Muhammad Abdurrahman, Amherst. GLSA. → pages 10

JI, HENG, JOEL NOTHMAN, BEN HACHEY, & RADU FLORIAN, 2014. Overview of TAC-KBP2015 tri-lingual entity discovery and linking. Procedural Text Analysis Conference (TAC2015). → pages 5

JÓNSDÓTTIR, MARGRÉT. 1989. Um *ir*- og *ar*-fleirtölu einkvæðra kvenkynsorða í íslensku. *Ísklenskt mál* 10–11.57–83. → pages 69

——. 1993. Um *ar*- og *ir*-fleirtölu karlkynsnafnorða í nútímaíslensku. *Ísklenskt mál* 15.77–98. → pages 69

KARTTUNEN, LAURI, & KENNETH R. BEESLEY. 2005. Twenty-five years of finite-state morphology. In *Inquiries into Words, a Festschrift for Kimmo Koskenniemi on his 60th Birthday*, ed. by Antti Arppe, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund, Anssi Yli-Jyrä, & Juno Tupakka, 71–83. Stanford: CSLI. → pages 24

KAWAHARA, SHIGETO. 2011. Experimental approaches in theoretical phonology. *The Blackwell Companion to Phonology* . → pages 8, 128

——, 2016. Psycholinguistic methodology in phonological research. Pre-print version for publication by Oxford Bibliography Online. → pages 8, 128

KNOKE, DAVID, & PETER BURKE. 1980. *Log-linear models*. Number 20 in Quantitative applications in the social sciences. Thousand Oaks: Sage. → pages 25

KRESS, BRUNO. 1982. *Isländische Grammatik*. Leipzig: Enzyklopädie Leipzig. → pages 61

KULLBACK, S., & R. A. LEIBLER. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22.79–86. → pages 118

LEGENDRE, G., Y. MIYATA, & P. SMOLENSKY. 1990. Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. In *Proceedings of the twelfth annual conference of the Cognitive Science Society*, 388–395. Cambridge: Lawrence Erlbaum. → pages 10

LEWIS, DAVID D. 1998. Naïve (bayes) at forty: The independence assumption in information retrieval. In *Machine learning: ECML-98*, 4–15. Springer. → pages 47, 85

MALOUF, ROB, & FARRELL ACKERMAN. 2010. Paradigm entropy as a measure of morphological simplicity. In *Proceedings from the Workshop on Morphological Complexity*, Harvard. HUP. → pages 3

MATTHEWS, PETER HUGOE. 1972. *Inflectional morphology: A theoretical study based on aspects of Latin verb conjugation*, volume 6. CUP Archive. → pages 5, 47

MCCARTHY, J.J., & A. PRINCE. 1993. Prosodic morphology I: constraint interaction and satisfaction. Technical report, Rutgers University. → pages 4, 9

MCMULLIN, KEVIN JAMES, 2016. *Tier-based locality in long-distance phonotactics: learnability and typology*. University of British Columbia dissertation. → pages 10

MOORE-CANTWELL, CLAIRE, & ROBERT STAUBS. 2014. Modeling morphological subgeneralizations. In *Proceedings of the Annual Meetings on Phonology*, volume 1. → pages 141

MORETON, ELLIOTT, & JOE PATER. 2012. Structure and substance in artificial-phonology learning, part I: Structure. *Language and linguistics compass* 6.686–701. → pages 10

MORTENSEN, DAVID, KARTIK GOYAL, SWABHA SWAYAMDIPTA, PATRICK LITTELL, ALEXA LITTLE, LORI LEVIN, & CHRIS DYER, in review. Unorthodox resource use allows rapid development of NER systems for 'low-resource' languages. → pages 5

MÜLLER, GEREON. 2005. Syncretism and iconicity in icelandic noun declensions: A distributed morphology approach. In *Yearbook of Morphology 2004*, ed. by Geert Booij & Jaap van Maarle, 229–271. → pages 61

PATER, JOE. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33.999–1035. → pages 10, 25

——, & ANNE-MICHELLE TESSIER. 2003. Phonotactic knowledge and the acquisition of alternations. In *Proceedings of the 15th International Congress on Phonetic Sciences*, Barcelona. Universitat Autònoma de Barcelona. → pages 4

PRINCE, ALAN, & PAUL SMOLENSKY. 2008. *Optimality Theory: Constraint interaction in generative grammar*. Hoboken: Wiley-Blackwell. → pages 4, 9, 25

PRZEPIÓRKOWSKI, ADAM, RAFAL GÓRSKI, MAREK ŁAZIŃSKI, & PIOTR PĘZIK. 2010. Recent developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, & Daniel Tapias, Valletta, Malta. European Language Resources Association (ELRA). → pages 91

R CORE TEAM, 2013. *R: A anguage and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. → pages 73, 78, 96, 117

ROBERT, CHRISTIAN. 2007. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. New York: Springer Verlag. → pages 117

SALONI, ZYGMUNT, WŁODZIMIERZ GRUSZCZYŃSKI, MARCIN WOLIŃSKI, & ROBERT WOŁOSZ. 2007. Grammatical dictionary of Polish. *Studies in Polish linguistics* 4.5–25. → pages 121

SCHEER, TOBIAS. 2012. Variation is in the lexicon: yer-based and epenthetic vowel-zero alternations in Polish. *Sound, structure and sense. Studies in memory of Edmund Gussmann.* 631–672. → pages 90

SCHENKER, ALEXANDER M. 1955. Gender categories in Polish. *Language* 31.402–408. → pages 87

SIMS, ANDREA D, 2006. *Minding the Gaps: inflectional defectiveness in a paradigmatic theory*. Ohio State University dissertation. → pages 138, 139

SPENCER, ANDREW. 1991. *Morphological theory: An introduction to word structure in generative grammar*. Hoboken: Wiley-Blackwell. → pages 5, 47

STUMP, GREGORY. 2001. *Inflectional morphology: a theory of paradigm structure*. Cambridge: Cambridge University Press. → pages 47

——, & RAPHAEL A FINKEL. 2013. *Morphological typology: From word to paradigm*, volume 138. Cambridge: Cambridge University Press. → pages 6, 55

TSUJIMURA, NATSUKO, & STUART DAVIS. 2011. A construction approach to innovative verbs in Japanese. *Cognitive Linguistics* 22.799–825. → pages 14

VAN ROSSUM, GUIDO, & FRED JR. DRAKE. 1995. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam. → pages 12

VANCE, TIMOTHY. 1991. A new experimental study of Japanese verb morphology. *Journal of Japanese linguistics* 13.145–156. → pages 128

VANCE, TIMOTHY J. 1987. *An introduction to Japanese phonology*. Albany, NY: State University of New York Press. → pages 128

WILSON, C. 2006. Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cognitive science* 30.945–982. → pages 24, 30, 45, 108

WOLIŃSKI, MARCIN, MARCIN MIŁKOWSKI, MACIEJ OGRODNICZUK, ADAM PRZEPIÓRKOWSKI, & ŁUKASZ SZAŁKIEWICZ. 2012. PoliMorf: a (not so) new open morphological dictionary for Polish. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, & Stelios Piperidis, Istanbul, Turkey. European Language Resources Association (ELRA). → pages 121, 122, 123

WYLLYS, RONALD E. 1981. Empirical and theoretical bases of Zipf's law. *Library trends* 30.53–64. → pages 6

ZWICKY, ARNOLD. 1985. How to describe inflection. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, ed. by Mary Niepokuj, Mary VanClay, Vassiliki Nikiforidou, & Deborah Feder, volume 11, Berkeley. University of California Press. → pages 5, 47

Appendix: supplementary materials

Icelandic experiment frame sentences

1. Jón safnar [DatPl].
2. Í gær bað Jón mig að gæta [GenSg] sem hann hafði fundið.
3. [NomSg] eru uppáhaldið hans Jóns.
4. Jón fann sex [AccPl] til viðbótar ídag.

Translations:

1. Jón collects [DatPl].
2. Yesterday, Jón asked me to take care of the [GenSg] he found.
3. [NomSg] are Jón's favorite.
4. Jón found six more [AccPl] today.

Icelandic experiment stimuli

In the Stem V column, *fu* indicates that the stem vowel is a front unrounded vowel, and *other* indicates that the stem vowel is not a front unrounded vowel.

Stem	DatPl Stem	AccPl	DatPl	GenSg	NomPl	Stem V
strep	strep	a	um	s	ar	fu
nep	nep	ir	um	ar	ir	fu
θret	θret	ar	um	ar	ar	fu
pet	pet	i	um	s	ir	fu
hrem	hrem	ir	um	ar	ir	fu
blen	blen	ar	um	ar	ar	fu
sken	sken	i	um	s	ir	fu
stem	stem	a	um	s	ar	fu
gleit	gleit	ar	um	ar	ar	fu
speit	speit	i	um	s	ir	fu
freip	freip	a	um	s	ar	fu
heip	heip	ir	um	ar	ir	fu
splein	splein	i	um	s	ir	fu
skeim	skeim	a	um	s	ar	fu
neim	neim	ir	um	ar	ir	fu
θein	θein	ar	um	ar	ar	fu
θap	θöp	a	um	s	ar	other
tvap	tvöp	ir	um	ar	ir	other
sprat	spröt	ar	um	ar	ar	other
vat	vöt	i	um	s	ir	other
flam	flöm	ir	um	ar	ir	other
tjan	tjön	ar	um	ar	ar	other
san	sön	i	um	s	ir	other
gam	göm	a	um	s	ar	other
mjót	mjót	ar	um	ar	ar	other
skrót	skrót	i	um	s	ir	other
klóp	klóp	a	um	s	ar	other
hnóp	hnóp	ir	um	ar	ir	other
θrón	θrón	i	um	s	ir	other
stróm	stróm	a	um	s	ar	other
kvóm	kvóm	ir	um	ar	ir	other
hón	hón	ar	um	ar	ar	other

Icelandic experiment demographic questionnaire (translation)

1. Is Icelandic your native language? [yes/no]
2. Have you taken a course in Icelandic grammar or linguistics at university? [yes/no]
3. What is your gender? [male/female/other/prefer not to respond]
4. When were you born? [ranges from 1937 to 1997/prefer not to respond]
5. If you have any questions or comments, please write them here. [text field]

Polish experiment frame sentences

1. W sklepie z zabawkami, Małgosia przyglądała się kolorowym [DatPl].
2. Naprawiłem ramię [GenSg], które odgryzł mój pies.
3. Na półce w pokoju Jasia stało dużo zakurzonych [GenPl].
4. [NomPl] to ulubione zabawki Jasia.

Translations:

1. At the toy store, Mary was looking at the colorful [DatPl].
2. I fixed the arm of the [GenSg] that my dog bit off.
3. On the shelf of Johnny's room were a lot of dusty [GenPl]
4. [NomPl] are Johnny's favorite toys.

Polish experiment stimuli

Stem	Gender	V-Stem	DatPl	GenSg	GenPl	NomPl
gęgiń	neut	gęgini	om	a	∅	a
kesiń	masc	kesini	om	a	y	e
ząziń	fem	zązini	om	y	∅	e
muriś	neut	murisi	om	a	∅	a
jubiś	masc	jubisi	om	a	y	e
ciliś	fem	cilisi	om	y	∅	e
cągiź	neut	cągizi	om	a	∅	a
nepiź	masc	nepizi	om	a	y	e
dąriź	fem	dąrizi	om	y	∅	e
myfić	neut	myfici	om	a	∅	a
żecić	masc	żecici	om	a	y	e
homić	fem	homici	om	y	∅	e
zązyń	neut	zązyni	om	a	∅	a
gezyń	masc	gezyni	om	a	y	e
comyń	fem	comyni	om	y	∅	e
cołyś	neut	cołysi	om	a	∅	a
wycyś	masc	wycysi	om	a	y	e
zakyś	fem	zakysi	om	y	∅	e
notyź	neut	notyzi	om	a	∅	a
zepyź	masc	zepyzi	om	a	y	e
nusyź	fem	nusyzi	om	y	∅	e
dubyć	neut	dubyci	om	a	∅	a
logyć	masc	logyci	om	a	y	e
mymyć	fem	mymyci	om	y	∅	e
lyzoń	neut	lyzoni	om	a	∅	a
zażoń	masc	zażoni	om	a	y	e
puchoń	fem	puchoni	om	y	∅	e
pecoś	neut	pecosi	om	a	∅	a
zyzoś	masc	zyzosi	om	a	y	e
cimoś	fem	cimosi	om	y	∅	e
żópoź	neut	żópozi	om	a	∅	a
wukoź	masc	wukozi	om	a	y	e
lagoź	fem	lagozi	om	y	∅	e
zęloć	neut	zęloci	om	a	∅	a
kęcoć	masc	kęcoci	om	a	y	e
ryboć	fem	ryboci	om	y	∅	e

(*continued on next page)*

Stem	Gender	V-Stem	DatPl	GenSg	GenPl	NomPl
cybań	neut	cybani	om	a	∅	a
cumań	masc	cumani	om	a	y	e
nimań	fem	nimani	om	y	∅	e
hulaś	neut	hulasi	om	a	∅	a
fątaś	masc	fątasi	om	a	y	e
łenaś	fem	łenasi	om	y	∅	e
rytaź	neut	rytazi	om	a	∅	a
gechaź	masc	gechazi	om	a	y	e
nąbaź	fem	nąbazi	om	y	∅	e
rucać	neut	rucaci	om	a	∅	a
pofać	masc	pofaci	om	a	y	e
łynać	fem	łynaci	om	y	e	∅

Polish experiment demographic questionnaire (translation)

1. Is Polish your native language? [yes/no]
2. Have you taken a course in Polish grammar or linguistics at university? [yes/no]
3. What is your gender? [male/female/other/prefer not to respond]
4. When were you born? [ranges from 1937 to 1997/prefer not to respond]
5. If you have any questions or comments, please write them here. [text field]