University of California
Los Angeles

# Productivity in Historical Linguistics

## Computational Perspectives on Word-Formation in Ancient Greek and Sanskrit

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Indo-European Studies

by

**Ryan Paul Sandell**

2015

© Copyright by
Ryan Paul Sandell
2015

Abstract of the Dissertation

# Productivity in Historical Linguistics

**Computational Perspectives on Word-Formation in Ancient Greek and Sanskrit**

by

**Ryan Paul Sandell**
Doctor of Philosophy in Indo-European Studies
University of California, Los Angeles, 2015
Professor Brent Harmon Vine, Chair

"Productivity" is a simultaneously familiar and fraught topic for all linguists. Every linguist believes that he can recognize it, yet grasping what properties characterize a "productive" process in opposition to a non-productive one is difficult. The historical linguist ought to be particularly desperate for a definition that can be operationalized – distinguishing archaism from innovation depends upon it. This dissertation is therefore first concerned with transforming "productivity" into an object that can be interrogated, and then seeking tools that can provide a useful characterization that object. On the explicitly diachronic side, I am concerned with how to measure, diagnose, and motivate changes in productivity. Corpora from two the oldest-attested Indo-European languages, Ancient Greek (here, mainly the *Iliad* and *Odyssey*, as well as the New Testament) and Vedic Sanskrit (here, mainly the *Ṛgveda*) will guide and define these explorations. Within the realm of morphology and word-formation especially, I will argue that concerns about productivity rightly take pride of place in diachronic discussion, and that those discussions become more meaningful when made precise and psychologically motivated. This increased precision offers hope of accounting for diverse linguistic phenomena for which the presence or absence of morphological structure is a crucial determinant. Overall, this work calls for the increased usage of corpus-based quantitative methods and computational modeling in the study of language change.

The dissertation of Ryan Paul Sandell is approved.

Stephanie J. Watkins

H. Craig Melchert

Bruce P. Hayes

Brent Harmon Vine, Committee Chair

University of California, Los Angeles

2015

*Per Chiara,*
*perchè nessuno potrebbe meritarlo di più*

# Table of Contents

# List of Figures

# List of Tables

* reconstructed form; marker of the head foot in Greek prosodic words
* [following] assured but unattested form
$^{X}$ hypothetical/incorrect form
$^{\star}$ ungrammatical; markedness constraint
# word boundary
- Morpheme Boundary
. Syllable Boundary
$<$ develops by sound change from
$>$ develops by sound change to
$<$ XX $>$ morphological or syntactic construction
$\rightarrow$ becomes in a synchronic derivation; approaches some value
$\infty$ infinity
$C$ normalizing constant
$\Delta$ change in
$E[X]$ estimate of $X$
$h_1$ corpus hapax legomena frequency
$\mathscr{I}$ pragmatic potentiality
$m$ token frequency index
μ mora
*N* token frequency
$n_1$ category hapax legomena frequency
$P(A)$ probability of event $A$
$P(A|B)$ conditional probability of event $A$ given event $B$
$\mathscr{P}$ Productivity in the strict sense
$S$ population type frequency
$\hat{S}$ estimate of population type frequency
σ syllable
*V* (sample) type frequency
*V*, hapax legomena frequency
$\mathscr{U}$ pragmatic usefulness
φ foot
ω word
$z$ Zipf rank

# Abbreviations

1 first person
2 second person
3 third person
AB Aitareya Brāhmaṇa
abl. ablative
acc. accusitive
act. active
aor. aorist
(Y/O)Av. (Younger/Old) Avestan
AV(Ś/P) *Atharvaveda* (Śaunaka/Paippalada)
Att.-Ion. Attic-Ionic
BĀU Bṛhadāraṇyaka Upaniṣad
C consonant
C. Classical
dat. dative
Dor. Doric
du. dual
E Epic
(f)ZM (finite) Zipf-Mandelbrot
gen. genitive
G(r)k. Greek
Hom. Homer(ic)
IE Indo-European
Il. *Iliad*
impv. imperative
inj. injunctive
inst. instrumental
JB Jaiminīya Brāhmaṇa
Lat. Latin
lim limit
loc. locative
log (natural) logarithm
MBL Memory-Based Learning
MBh Mahābhārata
MGL Minimal Generalization Learner/Learning
mid. middle
nom. nominative
NT New Testament
OCP Obligatory Contour Principle
Od. *Odyssey*
opt. optative
PDE present-day English
perf. perfect
(P)IE (Proto)-Indo-European
(P)IIr. (Proto)-Indo-Iranian
pl. plural
PPP past passive participle
pres. present
RV *Ṛgveda*(ic)
ŚB Śatapathā Brāhmaṇa
subj. subjunctive
sg. singular
Skt. Sanskrit
SR surface representation
V vowel; verb
Ved. Vedic
UR underlying representation
VGC vocabulary growth curve
WFR word formation rule

## Transcription and Transliteration

For convenience and easier legibility throughout the dissertation, I will largely cite Sanskrit forms using the standard International Alphabet of Sanskrit Transliteration (IAST) of the Devanāgarī script; symbols largely have IPA values, but please note the following conventions:

- Consonants with an underdot indicate retroflex place of articulation, but with features otherwise identical to corresponding dental consonants (e.g., *ṭ* = [ʈ]).
- Consonants with an underring are syllabic (e.g., *r̥* = [ɽ̩]).
- *c* and j are voiceless and voiced palatal stops (i.e., [c] and [ɟ]); *ñ* is a palatal nasal (i.e., [ɲ]); *y* is a palatal glide (i.e., [j]); *ś* is a voiceless palatal fricative (i.e., [ç]).
- Stops followed by *h* are aspirated in the case of voiceless stops (e.g., *th* = [tʰ]) or breathy voiced in the case of voiced stops (e.g., *dh* = [d̤]); *h* alone is a voiced glottal fricative (i.e., [ɦ]).
- A macron over a vowel indicates a long vowel (e.g., *ī* = [iː]).
- *a* is [ə], while *ā* is [aː].
- The vowels *e* and *o* are long (= [eː] and [oː]).

Greek forms throughout will usually be given both in Greek script and an accompanying phonemic transcription in IPA, though I resort to transcription alone in some places. The approximate phonetic values assumed are for 8th c. BCE Ionic Greek, unless otherwise indicated. English and other languages will usually be cited in standard orthography, unless the phonology of a form is strictly relevant.

# Acknowledgments

I have the habit, whenever I consult a dissertation, of sparing an extra minute or two to read the Acknowledgments. Approximately two pages is easily both the median and mode in length, but some outliers I have seen ramble into five pages, while others produce a terse solitary paragraph. Given my limited experience in acknowledgment-writing, I cannot say to have a good command of the art. I hope here to at least convey simple, accurate, and sincere gratitude.

To have even embarked on the process of writing this dissertation, let alone to have completed it, would have been unthinkable without the aid of my teachers and advisors, whose names appear on page iii. Brent, Stephanie, and Craig welcomed me into the Program in Indo-European Studies and saw fit to train me, even as a student of atypical background. Bruce gladly accepted students of Indo-European Studies into his courses, and it is mainly on account of what I learned in the classroom and in conversation with Bruce (as well as his colleagues in the Department of Linguistics, especially Kie Zuraw and Pam Munro), that I feel justified in regarding myself as a linguist. Indeed, it was in Bruce's course on Morphological Theory in the Spring of 2011 (now four years ago!) that I first read Hay and Baayen's "Parsing and Productivity" (Hay and Baayen 2002), which gave me the germ of the idea that would eventually sprout into this dissertation. Without the training in the arcana of Indo-European linguistics and familiarity with Homeric and Vedic philology gained from my hours with Brent, Stephanie, and Craig, however, to transform that germ into this product would have been impossible.

More important, I think, than the specific factual knowledge (*Wissenschaft*) gained from my time with these teachers, has been the formation and modeling for good scholarship (*Bildung*) that they each have provided. From Brent, I have learned indefatigable precision and attention to detail, as well as a wide-ranging curiosity that tries to unite disparate facts and methodologies. From Stephanie, I have obtained the model of the linguistic philologist, who tries never to forget to situate one's linguistic analyses against the background and constraints of the text. In Craig, I have seen that scholarly openness and unflagging intellectual honesty repay themselves many times over. From Bruce, I have learned to ask probing questions with a light heart, and to ask myself what the significance of my work is, so that I can invite others to understand it, too. All four of them generously allowed me to pursue research in directions that intrigued me. Now at the close of my years at UCLA, I wish only that I had availed myself of their counsel more often.

Naturally, none of my advisors named here is responsible for whatever imprecisions, inaccuracies, infelicities, and other errors of omission or commission that undoubtedly remain in this dissertation.

Conversation both in person and by e-mail with my colleagues and friends Chiara Bozzone, Andrew Byrd, David Goldstein, Dieter Gunkel, Jesse Lundquist, Tony Yates, and Sam Zukoff has influenced parts of this work in both subtle and obvious ways; I am happy to acknowledge their intellectual contributions to this work. To my sometime co-authors Andrew and Sam in particular, I give thanks for tolerating my occasionally interminable e-mails. I

thank Jesse for generously sharing his carefully curated data on Vedic *-ti-* abstracts that came to good use in Chapter 7. Conversation with Tony throughout the past year on problems of Indo-European prosody helped me to see better the possible connections between productivity and prosody, the result of which is Chapter 7.

Though I know them only as acquaintances or not at all, I wish to acknowledge Adam Albright, R. Harald Baayen, Jennifer Hay, Paul Kiparsky, and Donca Steriade, whose scholarship has influenced my thinking on crucial points. Without their work in areas of phonology, morphology, quantitative methods, and historical linguistics, this dissertation could not exist.

In Los Angeles, friendship from John Gluckman, Jesse and Mary Lundquist, and Tony and Sam Yates has been an invaluable source of moral support. Having reason to eat, hike, watch *Game of Thrones*, or have a whiskey at the end of the evening with them has prevented me from withdrawing into a lonely cavern. The long-time friendship of Anita Huizar-Hernandez, Joaquin Rios, and Nick Vann cannot be forgotten either. Anita's empathy concerning the academic job market during the past year has been particularly welcome.

The daily morning yowls of Mr. Jingles, the communal cat, were a signal to set myself to work on this project. He has been another welcome companion, and at times both welcome and unwelcome distraction, along the road of dissertating.

My family, especially my parents Paul and Carol, aunt Jean, grandmother Evelyn, sister Angie, and three-year old niece Maddie have all offered encouragement, advice, and good humor, distributed over the now nearly thirty years of my life, that help me to rest easily and to trust in my abilities. I do not doubt that they are proud of my accomplishments up to and including this work. I regret that my grandmother Evelyn Mae Bjornson (neé McArdle) did not live to see its completion; she would have sincerely enjoyed attending the doctoral hooding ceremony.

I cannot forget to mention here my teachers during my MA year at the Universiteit Leiden: Alexander Lubotsky, Michiel de Vaan, and Alwin Kloekhorst. They provided me with a basic grounding in Indo-European linguistics, without which my studies at UCLA would surely have failed. Furthermore, as guiding spirit behind the Leiden Summer School in Languages and Linguistics, I owe Alexander Lubotsky a great debt for (thereby indirectly) introducing me to Chiara Bozzone, whose friendship led me to seek a Fulbright Fellowship (from which the funding for my MA year I gladly acknowledge here) to study in Leiden.

And finally, for her boundless patience, support, sparkling intellect, wit, compassion, and love, I am profoundly grateful for Chiara. I have learned more from her, about all manner of things, during our seven years together, than from all others mentioned here. This dissertation is for her.

# Vita

| | |
|---|---|
| 2008 | BA Philosophy and BA Music *summa cum laude*, Barrett Honors College, Arizona State University, Tempe, Arizona |
| 2008–2009 | Fulbright Fellowship, Leiden, The Netherlands |
| 2009 | MA Indo-European Linguistics, Universiteit Leiden, Leiden, The Netherlands |
| 2010–2011 | Graduate Research Mentorship (under Prof. S. Jamison), UCLA |
| 2012–2013 | Teaching Assistant, Department of Linguistics, UCLA |
| 2013–2014 | Research Assistant, Center for Medieval and Renaissance Studies, UCLA |
| 2014–2015 | Dissertation Year Fellowship, UCLA |

# Publications and Presentations

& Andrew Miles Byrd. Under Contract (To Appear 2016). *The Life Cycle of a Sound Law: Szemerényi's Law as Study in the Diachrony of Phonological and Morphological Acquisition*. Amsterdam: John Benjamins.

& Sam Zukoff. Forthcoming. The Phonology of Morpheme Realization in the Germanic Strong Preterites. Proceedings of the 45th Annual North East Linguistics Society Conference.

& Andrew Miles Byrd. 2015. Extrametricality and Non-Local Compensatory Lengthening: The Case of Szemerényi's Law. 89$^{th}$ Annual Meeting of the Linguistic Society of America: 9 January 2015, Portland, Oregon.

& Sam Zukoff. 2014. A New Approach to the Origin of Germanic Strong Preterites. 45$^{th}$ Annual Meeting of the North East Linguistic Society: November 1 2014, Cambridge, Massachusetts

2014. On the Phonological Origin of Indo-European Long-Vowel ("Narten") Presents. 26$^{th}$ UCLA Indo-European Conference: October 25 2014, Los Angeles, California.

2014. Compensatory Lengthening in Vedic and the Outcomes of Proto-Indo-Iranian *[az] and *[až]. *Proceedings of the 25th UCLA Indo-European Conference*, S. Jamison, H.C. Melchert, and B. Vine (eds.). Bremen: Hempen Verlag.

2014. Sound Change or Grammar Change? Two Case Studies in Learner-Oriented Historical Phonology and Morphology. Indo-European Linguistics Colloquium between UCLA and the Ludwig-Maximillian Universität: 28 July 2014, Munich, Germany.

& Andrew Miles Byrd. 2014. In Defense of Szemerényi's Law. 30th East Coast Indo-European Conference: 6 June 2014, Blacksburg, Virginia.

2014. Perspectives on the Reconstruction of the Proto-Indo-European Accentual System. Kyōto-UCLA Indo-European Workshop. 26 March 2014, Kyōto, Japan.

2013. Compensatory Lengthening in Vedic and the Outcomes of Proto-Indo-Iranian *$az$ and *$až$. 224th Annual Meeting of the American Oriental Society: 17 March 2014, Phoenix, Arizona.

2012. Evidence for Acrostatic Presents in Old Irish?. *Proceedings of the 31st Harvard Celtic Colloquium*, D. Furchtgott, M. Holmberg, A. Joseph McMullen, and N. Sumner (eds). Cambridge, MA: Department of Celtic Languages and Literatures, Harvard University.

2012. Autopsy of a Morpheme: The Vedic 2.pl.act. Ending *-thana*. 222nd Annual Meeting of the American Oriental Society: 18 March 2012, Boston, Massachusetts.

2011. Evidence for Acrostatic Presents in Old Irish?. 31st Harvard Celtic Colloquium: 8 October 2011, Cambridge, Massachusetts.

2011. Reduplication and Grammaticalization in Vedic Sanskrit. 20th International Conference of Historical Linguistics: 28 July 2011, Ōsaka, Japan.

2011. The Morphophonology of Reduplicated Presents in Vedic and Indo-European. *Proceedings of the 22nd UCLA Indo-European Conference*, S. Jamison, H.C. Melchert, and B. Vine (eds.): 223–54. Bremen: Hempen Verlag.

2010. The Morphophonology of Reduplicated Presents in Vedic and Indo-European. 22nd UCLA Indo-European Conference: 6 November 2010, Los Angeles, California.

2009. The Evidence for 'Narten'-Roots in Greek. Leiden-Münster Indo-European Colloquium: 16 June 2009, Leiden, The Netherlands.

# Introduction

"Productivity" is a simultaneously familiar and fraught topic for all linguists. Every linguist believes that he can recognize it, yet grasping what properties characterize a "productive" process in opposition to a non-productive one is difficult. The historical linguist ought to be particularly desperate for a definition that can be operationalized – distinguishing archaism from innovation depends upon it. This dissertation is therefore first concerned with transforming "productivity" into an object that can be interrogated, and then seeking tools that can provide a useful characterization that object. On the explicitly diachronic side, I am concerned with how to measure, diagnose, and motivate changes in productivity. Corpora from two the oldest-attested Indo-European languages, Ancient Greek (here, mainly the *Iliad* and *Odyssey*, as well as the New Testament) and Vedic Sanskrit (here, mainly the *Ṛgveda*) will guide and define these explorations. Within the realm of morphology and word-formation especially, I will argue that concerns about productivity rightly take pride of place in diachronic discussion, and that those discussions become more meaningful when made precise and psychologically motivated. This increased precision offers hope of accounting for diverse linguistic phenomena for which the presence or absence of morphological structure is a crucial determinant. Overall, this work calls for the increased usage of corpus-based quantitative methods and computational modeling in the study of language change.

## Part I: Theoretical Bases

How is "productivity" logically to be understood and defined? Although the notion "productivity" is evidently important for historical linguists, Indo-Europeanists, and philologists alike, the conception of "productivity" in most works is pre-theoretical and not well-defined. As the historical survey in Bauer 2001: Ch. 2, and considerations under section 1.2.1 show, defining "productivity" is hardly straightforward. "Productivity" nevertheless has need of greater refinement and precision in order to be used meaningfully in historical linguistic research. To craft such a tool, and to make preliminary suggestions as to what functions it can serve, is the overall objective of Chapter 1. I argue that, while qualitative factors are transparently relevant, in many instances, in restricting the domain of application of a process, only an essentially quantitative approach is adequate for a full characterization of morphological productivity.

How is productivity to be measured? In Chaper 2, I introduce the corpus-based statistical methods originally developed in Baayen 1989 as the most promising basis for quantifying, and thus concretely discussing, morphological productivity. The fact that these measures rest on linguistic distributions in corpora makes them ideal for languages that exist only in the form of corpora, and for which no psycholinguistic data from native speakers can be obtained. Baayen's methods will be foundational to this entire study, but also deserve fair testing on languages and corpora that differ substantially from the modern English, Dutch, German, and Italian data to which those methods have been previously applied.

The question of a relationship between productivity and analogy also links pieces of Chapters 1 and 2. Departing from the idea that productivity and analogy are two sides of the same coin, I bring in the Minimal Generalization Learner (MGL) (Albright and Hayes 1999), originally developed as a model of analogy, as a means to model morphological productivity through morphological learning simulations.

Chapter 3 takes up questions of the psychological bases and connection of cognitive processes to productivity. How is productivity to be psycholinguistically interpreted? More pressingly, are the results that the quantitative measures provide demonstrably valid? I examine the correlations with attested experimental results, which substantially support some version of dual-route morphological processing. The way is then open to the investigation of how productivity may at least correlate with, if not indeed cause, other linguistic effects that depend upon the parsability of morphological structure.

Chapter 4 serves as a brief grammatical sketch of Ancient Greek and Vedic Sanskrit, naturally focused on the morphology. This chapter also briefly motivates the reasons for choosing the Homeric epics (the *Iliad* and *Odyssey*) and the *Ṛgveda* as the principal sources of data, and introduces the philological resources that substitute for deeply tagged morphological corpora of the Homer and the *Ṛgveda*.

## Part II: Case Studies

With toolkit in place, I delve into the data, to discern whether in fact that application of Baayenian and Albrightian methods can pass a "sanity check". Chapters 5 and 6 seek to apply these methods to the relatively easy cases of aorist formation in the Greek and Vedic corpora under study. Happily, all stars appear to align, and intuitively expected outcomes are externally confirmed. Comparison between the Greek and Vedic data, as well as extension of the Greek into the New Testament, allows for explicitly diachronic discussion of changes in productivity, and perhaps the novel reconstruction of productivity itself. I also consider the role of token frequency in helping forms belonging to non-productive categories to hold fast against the onslaught of productive neologisms. Results here align basically with earlier claims made by Bybee and Slobin (1982), and discussed for the history of English by Branchaw (2010): token frequency is a good predictor of morphological "stability", diachronically speaking.

Chapter 7 turns instead to the relationship between productivity and word-level prosody in Ancient Greek and Vedic. I argue in particular that some gaps in empirical coverage and arbitrary analyses under Kiparsky (2010)'s model of Vedic accentuation can be improved by explicit reference to synchronic morphological structure, which is not always true to etymological morphological structure. Quantitative exploration of productivity among several Greek and Vedic nominal categories in the light of work by Probert (2006b) allows me to develop the hypothesis that the non-productivity of a derivational process provides a necessary, but not sufficient, condition for synchronic loss of morphological structure, which in turn may reflect itself in a word's accentuation.

Where corpus-based measures of productivity appear to break down, and produce un-

interpretable or meaningless results, MGL may still step in and be of use. This is the case in Chapter 8, where tight morphophonological restrictions on productivity hold in the construction of Vedic perfect weak stems of the form $C_1eC_2$-. The project here strives to demonstrate that the presumed "analogical" expansion of this category cannot have occurred as traditionally described. Instead, MGL offers the means, in conjunction with new phonological observations, to establish a system wherein an analogy can properly "cascade" through generations of learners.

By way of conclusion, I summarize the concrete goals attained in the course of crafting this dissertation, and point to immediately related areas of further research. On the whole, this dissertation will serve to establish a method for measuring, interpreting, and applying the study of productivity in the practice of historical linguistics.

Part I

# Theoretical Bases of Morphological Productivity

# CHAPTER 1

# Morphological Productivity: Its Definition and Relevance for Historical Linguistics

To cite earlier remarks on the intractability of morphological productivity as a theoretical problem is *de rigueur* for any work that purports to concern itself with the problem:

- Aronoff (1976: 35): "Morphological productivity is one of the central mysteries of word-formation."
- di Sciullo and Williams (1987: 2): "[P]roductivity and listedness are not grammatical concepts."
- Bauer (1983: 62): "[P]roductivity remains one of the most contested areas in the study of word-formation."
- Mayerthaler (1981: 92): "'Produktivität' zählt zu den unklarsten Begriffen der Linguistik."[1]

Despite its seeming ungraspability, and assertions that its workings are part of "performance" rather than "competence", and hence outside the purview of linguistics proper (cf., e.g., Mohanen 1986: 56 or Sturtevant 1947: 122), the productivity of word-formational processes is not only a topic of frequent theoretical discussion, but indeed a theoretical prerequisite to any grammatical description that presumes that speakers have a capacity to somehow generate new words. Moreover, as I will emphasize under 1.3.1.1, the means to distinguish grammatical productivity from a paragrammatical creativity are available, and indeed, the fact that productivity can be made measurable will help to set it apart from creativity in language. Thus, unlike di Sciullo & Williams cited here, I hold that productivity in morphology is indeed susceptible to linguistic analysis.

At its heart, the problem of productivity seeks to answer questions of why and to what extent speakers generate or regard novel forms as acceptable. Productivity cuts through all aspects of grammar, from phonology to semantics: productive phonological processes decide what sequences of sounds are well-formed; productive inflectional morphology decides which morphosyntactic features do or do not require overt phonological expression; productive derivational morphology decides what kinds of novel lexical items are licensed; productive syntactic processes determine the hierarchical relations and linear orders that words assume; productive (compositional) semantics determines what interpretation is to

[1] "Productivity is among the most opaque concepts in linguistics."

be imposed upon some sequence of elements. Cases of categorical well- or ill-formedness in all of these domains are relatively easy to describe and understand: in Sanskrit, [+spread glottis] ("aspirated") stops are largely illicit within adjacent syllables (*buddhá-* 'awakened', **bhuddhá-*);[2] in Standard Italian, verbs obligatorily inflect for person, number, and tense through suffixes – pronouns or temporal adverbs cannot license omission of those affixes; in English, the adverb deriving suffix *-ly* accepts only adjectives as bases;[3] etc. Processes that apply irregularly or stochastically require a somewhat different outlook, though already Pāṇini describes "optional" rules or forms that optionally undergo rules, and the field of sociolinguistics has long employed regression techniques to analyze "variable rules" (already Labov 1969). With the appropriate combination of structural and social factors, to model the rate of word-final alveolar stop deletion in English (recently, Coetzee and Kawahara 2013) or the choice of linking segment in German compounds (Krott et al. 2007) is well within the realm of possibility.

Derivational morphology, however, as the remarks quoted at the outset should indicate, poses a particular problem. Namely, it is not immediately obvious why (1).a. seems entirely well-formed, (1).b. seems interpretable, albeit a bit strange, and (1).c., certainly when presented aurally (as I report anecdotally) elicits only confusion. Correspondingly, (1).a., though not listed in any major dictionaries of English, is easily found in natural texts, while (1).b. and (1).c, as far as I can easily determine, may never have seriously occurred. In (2), both a. and b. seem well-formed, while c. remains impossible.

(1) Abstract Noun Derivation in English – Native Base *orange*
   a. *orangeness*
   b. ?*orangeity*
   c. **orangeth*

(2) Abstract Noun Derivation in English – Latinate Base *perspicuous*
   a. *perspicuousness*
   b. *perspicuosity*[4]
   c. **perspicuousth*

While a difference in lexical base accounts for the difference in acceptability between (1).b. and (2).b., what makes the c. cases so disquieting escapes an immediate structural account – if attested abstract nouns that schoolchildren learn to segment with a suffix *-th* take adjectival bases, without any phonological restriction on the final segment of the base word,

[2] Stephanie Jamison reminds me that, while exceptionless in some morphological domains, e.g., reduplicated stems, a few forms in Sanskrit do contravene this constraint, e.g., *avabhr̥thá-* 'carrying away, removing'.

[3] One finds curiosities in the realm of branding, such as a website optimization startup named *Optimizely* (`www.optimizely.com`), evidently (wrenchingly, in my view) derived from the verb *optimize*.

[4] Cf. also the less transparent to the base *perspicuity*, which might block *perspicuosity*, just as *glory* might block ?*gloriosity*. Intuitively, I think that token frequency is decisive in these two cases: the higher token frequency *glory* more readily blocks ?*gloriosity* than lower token frequency *perspicuity* does *perspicuosity*.

what intuition so strongly excludes **orangeth*? There must, nevertheless, be a good linguistic motivation behind that fact that translations of Plato consistently render Gk. τραπεζότης [trapedzóteːs] as 'tableness' rather than 'tableth'.

This chapter will now develop the logical and theoretical basis for the dissertation. Before I can attempt to define morphological productivity, at least a few remarks on the place and role of morphology generally in grammar, and the sort of theoretical assumptions about morphology that I make, are necessary; this is section 1.1.[5] First, I will present the motivation behind a detailed study into the morphological productivity as part of language history (section 1.2). In particular, I confront the way in which the term "productive" is currently employed in historical linguistic (and especially Indo-Europeanist literature), and the uses to which "productivity" has been put. Section 1.3 is the heart of the chapter, wherein I establish a definition of morphological productivity, and point towards methods for measuring (detailed further in Chapter 2) and understanding productivity as a linguistic phenomenon. Then armed with a clear definition of productivity and awareness of its relation to other features of language I present some potential uses of productivity in the practice of historical linguistics under section 1.4.

## 1.1 Where's Morphology?

At the advent of generative grammar, morphology occupied a somewhat tenuous position in the conception of grammar as composed of "modules". In much research in the generative paradigm between roughly 1950 and 1975, morphology was exiled from the study of grammar proper, as it fell between the cracks of syntax and phonology; the current research program of Distributed Morphology (cf. Halle and Marantz 1993 and Halle and Marantz 1994 for the theoretical core, Noyer 2001 for a major empirical application) still maintains that morphology emerges from the interaction between semantic, syntactic, and phonological modules of grammar. In this vein, Borer (Forthcoming) argues strictly that all words, even forms with non-compositional semantics, are derivable through fundamentally syntactic operations. Conversely, the Lexicalist approach to morphology (first suggested in Chomsky 1970, classically developed in Jackendoff 1975 and Aronoff 1976) holds that "some members of major lexical categories (lexemes) are not derived by the same apparatus that derives sentences, but are inserted into lexical categories just as simple lexical items are" (Aronoff 2007: 804). As the nomenclature implies, the lexicalist view takes the lexicon, where all other listed idiosyncrasies are stored, to be the module where word-formation processes take place.[6] In the conception set forth in Aronoff 1976: 2–3 (building largely upon Halle 1973), inflectional morphology is essentially syntactic, whereas derivational morphology belongs to a morphological module of grammar. Here, word-formational rules (WFRs) produce lexical items, which can in turn be inserted into the syntax.[7]

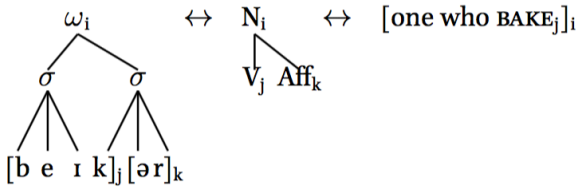
[5] At the risk of circular reasoning, I admit here that many facts about morphological productivity discussed below have in turn greatly informed my conception of what morphology is.

[6] For a summary of core works in lexicalist morphology from the 1970s, see Scalise 1984: 17–34.

[7] To distinguish between a "dictionary", in which all absolute, non-derivable irregularities, i.e., genuinely memorized forms, and the broader "lexicon", where word-formation processes take place, is also necessary.

For the purposes of this work, the development of Lexicalist morphology pursued in Booij 2010, Construction Morphology, is very convenient, because the formalism neatly ties together phonological forms with syntactic structures and semantic contents. Linguistic units that share all or some crucial set of those features, allowing for some variable elements, are "constructions", which can be readily identified from surface-level characteristics. At root, Booij follows Aronoff 1976: Ch. 2 in denying morpheme-based morphology because "the minimal linguistic sign is the word" (Booij 2010: 15), meaning that morphemes do not have a psychological representation independent of the constructional schemas (containing phonological, syntactic, and semantic properties) to which they are bound. Words such as *butcher* and *baker* instantiate a construction $<$ [ [x]$_j$ [ər]$_k$ ]$_{\omega i}$ $\leftrightarrow$ [[x]$_{Vj}$er$_k$]$_{Ni}$ $\leftrightarrow$ [PERSON WHO V$_j$-S HABITUALLY OR PROFESSIONALLY]$_i$ $>$. In example (3), the word shows *baker* co-indexed relations in the phonological, syntactic, and semantic components that relate their respective parts to one another:

(3) Lexical Representation of *baker* (after Booij 2010: Fig. 1.3)




Since Booij accepts accepts that the lexicon, morphology, and syntax all lie along a continuum, in which all patterns represent more or less abstract constructions, phrases like *I gave him the book* and *She wrote him a letter* are taken to instantiate a construction $<$ SUBJ V OBJ$_1$ OBJ$_2$ $>$.[8] This assumption is not a necessary prerequisite for the employ of constructional schemas in the morphological domain; McPherson (2014: 25–32) makes use of Booij's model, but explicitly states that "constructions exist only in the the case of idiosyncrasies, and whether at the word-level or phrase-level, I treat them as lexical morphology in the sense that they belong in the lexicon." For example, in Ancient Greek, verbal stems that terminate in a vowel often exhibit a phonological idiosyncrasy: the vowel lengthens when preceding either of two verbal derivational suffixes (aorist and future) that can only be analyzed phonologically as beginning with an /s/. Not all vowel stems are subject to participation this construction, thus making it doubly idiosyncratic: some stems must be lexically identified as

However, this distinction is often in practice ignored, and one must commonly interpret the expression "lexicalized" to mean "listed in the dictionary". Since the term "dictionarified" has no currency, I will likewise follow the practice of employing the terms "lexicalized" or "lexicalization" to refer to the treatment of (etymologically) morphologically complex words as morphological simplexes. The term "entrenchment" sometimes appears in psycholinguistic literature to refer to the same phenomenon, but also does not appear to be widely used.

[8] For further discussion and arguments for a lexicon–syntax continuum, see Jackendoff 2002 or Mos 2010. Jackendoff (pg. 52, fn. 4) suggests that, since Minimalist Syntax (Chomsky 1995) is basically a lexicalist theory of syntax, whose fundamental operation is to MERGE words whose lexical specifications permit for grammatical unification, it is not very distant conceptually from constructional syntax/morphology.

targets of this lengthening.[9]

(4) Vowel Lengthening Pattern in Sigmatic (/-s-/) Aorists
   a. pres. ἀλγε- [alge-] 'suffer pain' : aor. ἀλγησ- [algɛːs-]
   b. pres. ἀγορα- [agora-] 'speak' : aor. ἀγορᾱσ- [agoraːs-] (Hom. ἀγορησ- [agorɛːs-])
   c. pres. δουλο- [doːlo-] 'enslave, be a slave' : aor. δουλωσ- [doːlɔːs-]

These examples instantiate a specific subschema within a more general schema for Greek aorists built with the suffix /-s-/.

(5) General Schema: $<$ [ [...V]$_{Vj}$ [s]$_k$ ]$_i$ $\leftrightarrow$ [action of V$_j$ + PAST.PERFECTIVE$_k$]$_i$ $>$

(6) Particular Subschema: $<$ [ [...Vː]$_{Vj}$ [s]$_k$ ]$_i$ $\leftrightarrow$ [action of V$_j$ + PAST.PERFECTIVE$_k$]$_i$ $>$

I adopt no position here on the absolute relation between morphology and syntax, but I may employ constructional schemas as a representational convenience, even for cases where no phonological, syntactic, or semantic idiosyncrasies hold. I believe that this procedure is concordant with the methods of measuring morphological productivity described in Chapter 2, in providing a means of formally categorizing the members of a morphological category.

To my mind, the most serious question is whether morphological and syntactic "rules"/"schemas"/"constructions" obtain independent representations in memory, which can then be invoked independently of their instantiating exemplars, or whether the exemplars themselves continue to play a role the propagation of the processes they reflect. Albright (2008a), for instance, insists that symbolic morphophonological rules do indeed become exemplar-independent, while Daelemans and van den Bosch (2005) claim precisely the opposite, that no abstractions whatsoever are developed and stored, and the productive-seeming application of linguistic processes relies entirely upon stored exemplars. This latter view I think is too extreme, and, as discussion in Chapter 3 will make clear, difficult to reconcile with the psycholinguistic correlates of morphological productivity. On the likely need for symbolic or at least subsymbolic (connectionist) approaches, in which some degree of abstraction from exemplars is possible, see Booij 2010: 88–92, 258–9.

In sum, I will adopt throughout a Lexicalist approach to morphology, which understands that derivational processes, at least, take place in the lexicon (in the broad sense; cf. fn. 7 above). Morphological processes may be variously represented using output-oriented constructional schemas as in Booij 2010, or through input-based morphophonological rules as in Albright 2002b. I will subsequently argue that models of morphological production and processing that eschew any form of abstraction pose worrisome conceptual difficulties, and perhaps fail to adequately identify productive processes.

[9] This phenomenon will be revisited in Chapter 5.

## 1.2 The Problem of "Productivity" in Historical Linguistics

The notion of "productivity" is far from an unfamiliar one in the practice of historical linguistics. In particular, awareness of whether (or not) a given type of formation was "productive" at some given time, so as to have generated some particular form, is often crucial to arguments about the history of a form, or the history of a morphological category. The basic reasoning, which is often employed in practice (though I have yet to encounter a very explicit formulation), runs as follows:

1. Form/Construction X exhibits a pattern (call it type Y) that is unknown/rare in other forms/constructions in the language.
2. Therefore, forms/constructions of type Y are non-productive – type Y is non-productive.
3. Forms/Constructions belonging to non-productive types are chronologically older than forms/constructions belonging to productive types.
4. Form/Construction X is old.

In effect, "rarity" would appear to be the standard by which productivity is assessed – but on what (presumably quantitative) basis is "rarity" defined? The genuine problem is how to determine, in a non-circular fashion, which processes are synchronically productive and which are not. This problem is worsened for archaic corpus languages, where data from broad chronological periods are sometimes lumped together, despite the fact that substantial changes in grammars may distinguish different periods.

Let us consider a specific example in greater detail. While discussing denominal adjectives derived with the suffix *-ra-* in Sanskrit, Wackernagel (1905: 59–61) observes that "nur in Trümmerstücken ist die alte Regel bewahrt, daß das adjectivsuffix *-ra-* im Vorderglied von Komposita durch *-i-* ersetzt wird," and further remarks that "im A[lt]i[indischen] dieses kompositionelle *-i-* früh aufgehört **lebendig** zu sein."[10] Two methodological questions immediately come to the fore:

1. on what basis has Wackernagel (following Caland 1892) concluded that the use of this "compositional *-i-*" reflects an "old rule"?
2. how has Wackernagel discerned that the use of this "compositional *-i-*" "ceased to be productive"?

Wackernagel's reasoning is straightforward: forms containing "compositional *-i-*" are found mainly in the earliest Vedic text, the *Ṛgveda*; parallel formations without the "compositional *-i-*" are also found in the RV, and in younger Vedic texts; identical forms containing "compositional *-i-*" are also known in Avestan. Avestan cognates establish a Proto-Indo-Iranian

[10] "...only in fragments is the old rule preserved, that the adjectival suffix *-ra-* is replaced by *-i-* in the first member of compounds...in Old Indic, this compositional *-i-* early ceased to be living."

date for some words, thus implying that the forms are "old"; the fact that previously unattested forms with "compositional *-i-*" are rare in younger texts, and that other forms substitute forms that previously exhibited it, together imply that the word-formational rule has been given up. This reasoning is philologically solid: (partial) word equations (e.g., Skt. *á-kravi-hasta* 'not having bloody hands' : Av. *xruui-* 'bloody') allow for ready reconstructibility (PIIr. */krau̯H-i-/); the replacement of forms attested in older texts by other forms in younger texts allows for the conclusion that whatever process produced the forms largely restricted to the older texts is no longer operational in the younger texts. A diachronic change in morphology seems to be at work. The capacity of a morphological process not merely to generate new words, but even to maintain existing words that instantiate it, has diminished.

Yet, a perilous grey zone exists between the old stage (here, reconstructed Proto-Indo-Iranian) and the attested situation, at least with respect to the reconstructibility of specific forms. If the use of "compositional *-i-*" in *-ra-* forms was, at some point in the prehistory of both Vedic and Avestan, "lebendig", how is the independent formation of *-kravi-* and *xruui-* in Sanskrit and Avestan respectively to be excluded? The crucial missing piece of information is the DEGREE OF PRODUCTIVITY of this type of formation, synchronically, in the relevant periods of the languages. The formation's diachronic path in Sanskrit demonstrates that "compositional *-i-*" becomes unproductive, but since it still exists in the oldest texts, it is might not be totally unproductive at that time.[11] To know the DEGREE OF PRODUCTIVITY of "compositional *-i-*" would go a long way towards being able to assess the likelihood that *-kravi-* and *xruui-* are independent formations (or not). A means of reconstructing that DEGREE OF PRODUCTIVITY in the last common stage of Sanskrit and Avestan would in turn allow one to see directly the changes in productivity that a category has undergone diachronically.

Another serious concern is that Wackernagel does not describe or try to uncover the factors that allowed "compositional *-i-*" to be productive in the first place, nor the factors that caused it to become unproductive. Discussion in section 1.3 and in Chapter 2 considers the fact that factors both internal and external to a linguistic system have a role in a morphological process' DEGREE OF PRODUCTIVITY, but the linguist's explicit task is at least to point out the relevant structural factors. Thus, while Wackernagel leaves us secure in thinking that the morphological pattern of *-ra-* replacement by *-i-* in compounds did at one time generate forms, but later ceased to, and was replaced by a different pattern ("...drang einerseits das *-ra-* des Simplex in die Zusammensetzung"[12]), we do not know the reasons that underlie this change in word-formation, nor the extent to which the exemplars of the pattern that we have are properly old or new.

The point is, then, that we require further methods to aid in the study of diachronic morphology. Specifically, we require a better grasp of the notion "productivity" and of what an understanding of that notion can offer to the practice of Historical Linguistics. I delay the problem of what an adequate measure of morphological productivity can resemble until section 1.3. Before that, I wish to examine the sort of uses for "productivity" that Indo-European

[11] Thus whether "compositional *-i-*" was totally unproductive or not depends upon whether or all forms that instantiated it were lexicalized, or whether such forms instantiated a morphological pattern that could be extended.

[12] "...on the one hand, the *-ra-* of the simplex came into the compound."

linguistics has found, and also to discuss ways in which scholars have explicitly studied the productivity of word-formation diachronically.

### 1.2.1 "Productivity" in Indo-European Linguistics

The Indo-Europeanist has a particularly difficult task in trying to discuss linguistic productivity. Not only does the Indo-Europeanist's interest extend to reconstructed languages, for which, obviously, no direct testimony is available, but for the languages that form the basis of the field, knowledge is incomplete, and the lack of active native competence of the languages constrains (or should constrain) confidence in intuitions about productivity. While some trends and changes in the diachronic word formation may be so clear that even an informal characterization would be adequate, in the many more subtle cases, the philologist should worry that a non-native intuition based solely on written materials might go astray;[13] in addition, given a "Sprachgefühl" based on texts from different periods, that intuition may further deviate from the intuitions of any speakers that ever existed.[14] Panagl (1982: 228) thus rightly wonders "ob wir auf eine Beurteilung der Produktivität von Wortbildungstypen in Corpussprachen oder gar auf rekonstruierten Sprachstufen gänzlich verzichten müssen."[15]. Panagl himself is not pessimistic about the possibility of properly reckoning with productivity in corpus languages, but the indirect approaches that he proposes seem to have found little audience. As the succeeding sections show, the problem of how to reckon with productivity in Indo-European linguistics is very much alive.

### 1.2.2 Archaism versus Neubildung

As linguists who employ reconstruction as part of their regular practice well know, not every form can blindly be phonologically projected back to a proto-language; many new lexemes, through one means or another, may have entered a language well after the dissolution of its proto-language. Indeed, the term *Transponat* has developed precisely to refer to phonological reconstructions that probably were not words of the proto-language. All too often, however, whether a reconstruction can have reality, or is a mere Transponat, is uncertain. Both of the two major collections on the reconstruction of Indo-European verbal and nominal formations, Rix et al. (2001) (*Lexikon der indogermanischen Verben*: LIV) and Wodtko et al. (2008) (*Nomina im indogermanischen Lexikon*: NIL), expresses concerns about the effects of productive categories in the Indo-European daughter languages when attempting to judge the pedigree of a form.

In the LIV, many lemmata contain forms classed as "Neubildungen" ("recent forma-

[13] Indeed, even for modern languages, Coppieters (1987) has shown that the intuitions of non-native speakers differ substantially from the intuitions of native speakers.

[14] This problem is not universal: in the case of some corpora (e.g., Gothic), the sample is limited to a short period, and perhaps even a single speaker (though these circumstances may present different issues; the Gothic corpus in particular faces the issue of being translation literature).

[15] "whether we must altogether give up on an assessment of productivity of word-formation types in corpus languages or in reconstructed stages of languages."

tions"), which the editors define principally as following "einzelsprachlich produktiven Bildungsregeln" (pg. 13). That the editors of the LIV are concerned with distinguishing inherited formations from formations that were most likely created at the *einzelsprachlich* level is apparent, but the means for in fact executing such a division are more fuzzy. The editors of the LIV apparently presume to know what the "einzelsprachlich produktiven Bildungsregeln" are, but if they possess a method for determining them, beyond intuition or common sense, they do not explicitly say so. Moreover, they editors suggest that they need explain their decisions only "wo die Entscheidung nicht evident ist" – some productive processes are presumed to be self-evident. Finally, the editors admission that the categorization of forms could be revised "mit neuem Gründen" seems to call out for a more decisive means of separating inherited forms from innovative forms.

In a yet more cautious, or even pessimistic vein, the editors of the NIL worry that the distinction between an Indo-European form and form created in a daughter language "ist freilich oft nicht feststellbar" (pg. XV). Distinguishing clear renewals that follow obviously Indo-European patterns of word-formation from genuine inheritances bothers the editors of the NIL, but they see no effective means of escaping the conundrum. Perhaps the best solution for this problem is a probabilistic solution: how likely is it that a particular form was built productively in a given language at a given time? With the means to measure productivity, this sort of probabilistic evaluation of archaism versus Neubildung may be possible: in raw terms, the less productive a category is, the more likely that a given form belonging to that category predates the time at which productivity of the category is assessed. Internal to categories, are either general or category-specific factors, which could serve as faithful indicators of a given lexical item's age, available?

What the remarks on productivity from both the LIV and NIL make entirely clear is that, while "productivity" already rightfully plays an essential role in the study of word-formation in the oldest Indo-European languages, methods to develop satisfying and falsifiable claims concerning that "productivity" are wanting.

### 1.2.3 Use of the Term 'Productive': Rau 2009

Remarks taken from Rau 2009 will serve as a foil, in order to illustrate the extent to which the terms 'productive' and 'productivity' fill discussion on Indo-European morphology, even without a very precise understanding of productivity. I aim here to deconstruct Rau's usage of the term in order to gain an understanding of what a solid work on Indo-European morphology takes as evidence for productivity, and how knowledge of productivity is put to use in making linguistic claims. Moreover, since the book is precisely concerned with derivational morphology in PIE and its daughters, the problem of productivity is ever-present.[16]

Above all, Rau's judgments concerning productivity, and I suspect many of the judgments in other literature to which he refers, rest primarily upon type frequency. For example, the statement that factitives built with the suffix *-nu* "ha[ve] become productive in

[16] I counted 46 uses of the words '(un)productive' and '(non-/un)productivity' in 175 pages of body text and footnotes in Rau 2009. Contrast 30 occurrences of 'nominal' (a word occurring in the title) in the same span.

denominative function in Hittite" appears to rest on the fact that a number of such verbs in *-nu* indeed exist, combined with the fact that the affix is often applied to bases that are not likely to have an Indo-European origin. However, type frequency alone is not an adequate measure of productivity.[17] Furthermore, the frequencies of the *nu*-factitives would require comparison to the frequency distributions of other morphological categories in order to establish whether they really exhibit the profile of a productive category.[18]

One overarching feature that connects all of Rau's applications of 'productive' is that assertions of the form "X is productive" are often presented as self-evident. At best, Rau offers a reference. For instance, on pg. 56, he points to Wackernagel and Debrunner 1954: 754 regarding the productivity of *man*-stems in Indo-Iranian, or to Leskien 1891: 244 ff. on pg. 170 for the productivity of deverbative *u*-stem adjectives in Lithuanian. A reference, however, is more the exception than the rule. In the main, one finds little explicit reasoning, even of the sort that the Wackernagel gave concerning "kompositionelle *-i-*" (cf. above), to support a claim concerning productivity. For example, on pg. 74, we read that "amphikinetic" *s*-stems are "especially productive in Latin", though only three examples are cited (*rubor*, *tenor*, *furor*).

Occasionally, the forward-looking diachronic reasoning of the Wackernagel type occurs, e.g., on pg. 143, where Rau supports the claim that the suffix *-áya-* "is productive in deverbative function in Indo-Iranian" on the basis of the fact that "many of the formations listed here have replaced older factitive present types", e.g., the AV has the *-áya-* form *śobhayati* against Class VI present *śumbháti* 'beautifies' in the RV, and likewise AV *pūráyati* versus RV Class IX *pr̥ṇā́ti* 'fills'. Strictly speaking, however, this argument does not establish any claim about the productivity of *-áya-* in either the RV or the AV, only that *-áya-* has encroached upon the domain of usage for other present-forming suffixes sometime between the composition of certain RV verses and certain AV verses. Moreover, this pattern of replacement, which demonstrates that *-áya-* and other formations were in competition to some extent, could perhaps be not so much indicative of the (increasing) productivity of *-áya-*, but rather the non-productivity of the other formations (though at a practical level, this distinction amounts to the same thing).

More problematic is Rau's phrase "X enjoyed some/a period of productivity" which seems to imply that a type of formation was once productive at a particular phase in the history of a language, thus generating a few attested types, but is no longer. On pg. 140, fn. 48, Rau describes such a situation for the suffix */-d$^{h}$e/o-/ in the history of Greek, which would be responsible for verbs such as Grk. πλήθω [plɛ́ːtʰɔː] 'be(come) full' and βρῑ́θω [bríːtʰɔː] 'be weighed down'. The reasoning here is that if a type of formation lacks a thorough compar-

[17] See also Baayen (1989: 27–41) for devastating critiques of quantitative measures of productivity based solely on type frequency. Attempts to measure productivity based purely on type counts as recorded in standard dictionaries (i.e., based on neologisms explicitly recognized and catalogued by lexicographers) are patently inadequate, because they overlook large numbers of the rare and unique words that are precisely reflective of productivity; cf. the discussion and criticism of a dictionary-based study, Cannon 1988, in Baayen and Renouf 1996. This issue does not emerge for the corpus languages with which I am concerned, because usually every lexical item occurring in texts in that language is documented.

[18] See Chapters 2 and 5 for an account of how this can be done.

ative basis in related languages, and does not give the impression of being productive at an attested phase of the language, then it must have been productive at some preceding stage, in order to account for the attested types. Such circumstances seem genuinely difficult to assess.

The ultimate point here is not that Rau's specific claims are wrong – in large part, my own intuitions concerning relevant languages leads me to believe that many of his claims are right – but that, because productivity is an undefined entity in his work, to meaningfully evaluate or to test these assorted claims is difficult. All of these remarks do, however, further serve to substantiate the centrality of "productivity" in diachronic morphological research, and underscore the need for the need for a better toolkit.

### 1.2.4 A Qualitative Diachronic Approach to Productivity: Gardani 2013

One possible toolkit, based in the framework of Natural Morphology (Dressler 1985, Dressler 1987, Dressler 1999), is developed by Gardani (2013) as a means of accounting for the historical trajectory of nominal class membership between Latin and Old Italian (800–1400 CE). Gardani's criteria for the analysis of productivity are explicitly qualitative, and assessed on the behavior of nominal inflection classes and their subhierarchies. Insofar as the project is concerned with the productivity of inflectional classes and their specific patterns, the problems studied are closer in nature to work on analogy in inflectional paradigms (compare most explicitly here Ch. 4 of Albright 2002b on Latin) than to most other work on morphological productivity, which centers the discussion of productivity on derivational morphology (e.g., Aronoff 1976: Ch. 3, Baayen 1989, Bolozky 1999, Plag 1999).

Crucially, Gardani (pp. 19–20) regards both type and token frequency as merely derivative of productivity; yet, simultaneously, he claims that "productivity does not give clues as to the probability with which a new word or new form can be produced or accepted." Granting that productivity as an independent property indeed determines frequency, and that frequency can measure production probabilities, then logically, productivity should directly correspond to production probabilities. Gardani must then be (implicitly) rejecting the second premise (that actually attested frequency of occurrence can measure future production probabilities), but this is demonstrably wrong – empirically, at the level of lexical items, frequency of occurrence in one sample of a language is a very good predictor of occurrence in another sample – and moreover is inconsistent with psychological evidence (see further Chapter 3).

Setting aside any possible criterion grounded in frequency, Gardani proposes instead to measure productivity on the grounds of which inflectional classes accept new members (the "openness" of those classes). Gardani establishes four characteristics intended to measure the integrability of a lexical item into the existing inflectional system: SIMILARITY, FOREIGNNESS, NEWNESS, and INFLUENCE OF DERIVATION. Thus, words that are not similar to any known words, are foreign, and are new, can fall into only the most productive classes. I reproduce Gardani's scale for inflectional productivity here (cf. Gardani 2013: 46–8, 70):[19]

[19]The same criteria, Gardani says, hold *mutatis mutandis* for establishing the productivity of derivational

(7) Qualitative Productivity Scale

a. Inflection class assignment to loanwords with incompatible properties – high productivity
b. Inflection class assignment to conversions – mid-high productivity
c. Inflection class shift – mid-low productivity
d. Inflection class assignment to loanwords with incompatible properties under the influence of a productive derivational affix of the recipient language – low productivity
e. Inflection class assignment to loanwords with compatible properties – low productivity

Although this scale permits of a gradual rather than absolute interpretation of productivity (see further section 1.3.1), because it is not continuously valued, the comparison of productivity between classes that fall into the same discrete gradations is not possible. For instance, the feminine 1st and masculine/neuter 2nd declension nouns in Latin, and their most direct continuants in Old Italian (fem. nouns like sg. [kasa] : pl. [kase], masc. nouns like sg. [libro] : pl. [libri]) are rated as highly productive throughout all chronological periods that Gardani examines, and so the extent to which one or the other of those two classes is more productive is not measurable. Furthermore, the fact the above criteria depend heavily on the availability of loanwords or conversion processes as diagnostics limits their applicability to languages and chronological periods that take in loanwords and apply conversion as a morphological process. Hence, while this productivity scale may succeed reasonably well in the assessment of Latin and Italian inflectional material, how it could be adapted as an analytical tool to the circumstances of any whatsoever language is far from evident.

I must therefore conclude that the qualitative approach to morphological productivity briefly surveyed here exhibits some severe limitations. Perhaps most importantly, this overall theory of productivity does would appear to entirely lack a plausible theory of learnability: how is it, in Gardani's model, that native speakers obtain the competence to judge processes as unproductive or not, or which can predict the behavior of speakers? As I attempt to better define "productivity" in the succeeding section, I hope to show that any attempt to frame and describe productivity independent of a quantitative element may want for sufficient subtlety and precision. While qualitative factors may in part underlie the productivity of a process, the potential number of such factors and their interactions in linguistic phenomena, reflecting the diversity of elements that human cognition may track (cf. Barth and Kapatsinski 2015: 2–7) is likely too great to handle manually.

processes as well.

## 1.3 Theoretical Approaches to Morphological Productivity

### 1.3.1 Defining Productivity

As mentioned at the outset of this chapter, the question of productivity is a question of the grammaticality of words, parallel to the grammaticality of utterances. In the same fashion in which several words can come together with an infelicitous result, a word itself might be infelicitous. Formulated pre-theoretically, the question is of the form: "why is *steepness* more well formed than **stepth* (which is in turn more well formed than, e.g., **perspicuousth*)?" In essence, the problem is how words organize themselves into morphological patterns so that a speaker concludes that he should produce one form over another, particularly in the case where he has not previously perceived or produced a word that could fill the desired conceptual role. Indeed, native speakers, linguists, and philologists can arrive at generalizations such as "the suffix *-ness* is more productive than the suffix *-th*" or "Class I presents (i.e., present stems formed with the suffix *-a-*) in Vedic are more productive than Class III presents (i.e., present stems formed with partial reduplication)." All of these intuitions, insofar as they are reliable and accurate, must have a basis in some psycholinguistic reality. The first step is to try to make the notion "productive" more concrete and theoretically meaningful, so that a statement "X is productive" has an unambiguous interpretation.

A first-order concern is whether productivity is a quantitative or a qualitative notion. If the latter, productivity could be spoken of in terms of a binary feature (or maybe some bundle thereof) [+/- productive] that a word-formation rule or construction could have. If productivity is a quantitative notion, then it could be a wholly scalar notion and gradual, with "non-productive" (productivity = $0$ = "the formation never occurs in a word") and "fully productive" (productivity = $1$ = "the formation occurs in every word") being merely endpoints. Perhaps better would be to say that a productivity of $0$ indicates no further potential domain for application, i.e., there are no possible unrealized inputs to the WFR or construction, while a productivity of $1$ would indicate a completely open and as yet unrealized domain of productivity.

As a starting point for discussion, I take three definitions of productivity: 1) two very general pre-theoretical definitions; 2) a much-cited qualitative definition (cf. Plag 1999: 13, Baayen and Lieber 1991, Cowie and Dalton-Puffer 2002: 412–4); 3) a definition that makes productivity explicitly scalar and thus quantitative:

1. (a) Plag (1999: 6): "Productivity is generally loosely defined as the possibility to coin new complex words according to the word-formation rules of a given language."

   (b) Bauer (1983: 100): "A morphological process can be said to be more or less productive according to the number of new words which it is used to form."

2. Schultink (1961: 113): "Onder produktiviteit als morfologisch fenomeen verstaan we dan de voor taalgebruikers bestaande mogelijkheid door middel van het morfologisch procédé dat aan de vorm-betekniscorrespondentie van sommige hun bekende woorden ten grondslag ligt, onopzettelijk een in principe niet telbaar aantal nieuwe for-

maties te vormen."[20]

3. Bolinger (1948: 18): Productivity is the "statistical readiness with which an element enters into new combinations."

Bauer and Plag's remarks serve nicely, in that they encompasses several key definitional issues, "possibility" and "new" among them. From the outset, morphological productivity evidently requires "word-formation rules" or a "morphological process" in order to generate forms (to "form" (Bauer) or "coin" (Plag) them), just as syntactic productivity involves the formation of larger phrases and utterances on the basis of syntactic "rules" (cf. Chomsky 1965: 6).[21] The further questions to be pursued from Plag's and Bauer's definitions are:

- a) why should productivity involve only "complex" words?
- b) what counts as a "new" word?
- c) what is a "possible" word, or what allows for the "possibility" of generating new words (through whatever means)?

a) That the study of morphological productivity should involve only complex words is really a practical restriction. As a matter of fact, morphologically simplex words enter a language very rarely (probably most commonly as loanwords), and we will subsequently see that the degree of productivity of simplex words can serve as a baseline and comparative measure of what is *not* productive. The reason for the fact that new simplex words appear infrequently is intuitive: given that words are signs, a person can only exceptionally craft a totally new sign and expect to achieve a pragmatic communicative objective. At the same time, not all ostensibly simplex words are necessarily so, as in cases of zero derivation, and in such cases, which member is non-derived is not always apparent (or interpretations may differ between speakers; cf. (Plag 1999: 219–25)). The ultimate point here is that probably only morphologically complex words, in the sense that a linguist, at least, can make some kind of morphemic analysis, can ever have a morphological pattern to propagate.

b) The issue of what counts as a "new" word, or neologism, is not at all an absolute matter. From the perspective of an individual speaker, a "new" word is simply a word that the speaker has never before encountered; words of extremely low frequency might even be perceived as "new" by the same speaker on multiple occasions.[22] Hence, a word might be familiar to other speakers of a given language, or have been recorded in a dictionary by lexicographers,

[20] Translation after van Marle (1985: 101): "We understand productivity as a morphological phenomenon to be the possibility for language users to coin, unintentionally, a number of formations that are in principle uncountable, by means of the morphological process that underlies the form-meaning correspondence of some words already known to them."

[21] That morphological productivity and syntactic productivity do not really differ, in that they both instantiate linguistic productivity, is a position held in Beard 1977: 332–4, Bauer 1983: 72–4, Corbin 1980, Mos 2010, and Zeldes 2012.

[22] A "novel" encounter with the same word on multiple occasions presumes that the word might be so infrequent that all traces of its existence have disappeared from a speaker's memory.

yet still may be entirely new to a speaker. On the other hand, absolute neologisms, words that no speaker has ever before produced or perceived, must perforce exist. In the case of morphological categories that readily generate new members, the number of absolute neologisms will be higher. The importance of neologisms, or statistically rare words that stand in for neologisms, will become further evident under section 2.2.1.[23]

c) The trickiest issue that in defining productivity involves "possibility" or "potentiality" in word-formation. For discussing this issue, I turn to the second definition, Schultink's, which introduces two further terms: "uncountable" and "unintentional". These issues occupy the following subsection.

#### 1.3.1.1 "Productivity" and "Potentiality" versus "Creativity"

Already de Saussure (1983: 227), and later Aronoff (1980: 163), worried about the distinction between "potential" and "actual" words; in Aronoff's view, the lexicon determines what words are actual, all others being potential. Strictly speaking, potential words, for Aronoff, are only those words that a WORD-FORMATION RULE of the speaker can generate, or like simplex words, require explanation or substantial context for understanding. Here, Schultink's notion that a productively formed word be "onopzettelijk" ("unintentional") comes into play. Following Baayen and Lieber (1991: 808), "unintentionally" coined words, i.e., words formed through productive processes, "will go unnoticed" whereas intentional coinages, which employ unproductive or otherwise non-existent word-formational properties "will be used to shock, amuse, or achieve some other intentional effect."[24]

Yet, formations that were probably truly creative in nature may come to have a sufficient number of exemplars so as to engender a grammatically productive processes. The English affix -(*a*)*thon*, used to denote an event, especially a competitive or fund-raising function, that carries on for a lengthy period, is in origin "creatively" (OED 2013: s.v. *-athon*: "barbarously") extracted from *marathon*. *-athon* now productively builds new forms, behaving more like a normal affix, insofar as new coinages that employ it are readily comprehensible; the current OED contains no less than eleven entries containing -(*a*)*thon*.[25] Indeed, words containing this affix, because they now seem to be happily "potential" words, may escape the notice of lexicographers, who may judge that the form must have already been recorded (cf. Baayen and Renouf 1996: 74–5). The ultimate point is that "creative" coinages are not "potential" in the grammatical sense. Although "creative" coinages must be a reflection of some sort of linguistic capacity, I would argue that they are at best paragrammatical.

[23] Bauer (2001: 38–9) distinguishes nonce-words from neologisms: the former genuinely occur but once in a language, and never obtain purchase among speakers, whereas neologisms become "part of the norm of the language, and thus part of the brief of a lexicographer." I will continue to use "neologism" to mean a new coinage, whether intentional or not; see further the following subsection.

[24] The difficulty of attempting to fit "creative" coinages into grammar must be what Sturtevant (1947: 122) intends in writing: "coining words, like writing books, is a function of artists…We linguists need not attempt the hopeless task of classifying the inventions of advertisers, philosophers, and – linguists."

[25] Lehrer (2007) refers to cases like these as "splinters" (sometimes also referred to as "voguish affixes"), and identifies *-gate*, -(*a*)*holic* and -(*a*)*thon* as having achieved the status of independent morphemes in Present-Day English.

To distinguish between productive and creative word-formation is all the more difficult with corpus languages, for which knowledge of the grammatical properties of the language is limited to the corpus itself, and the completeness of the historical record constrains awareness of non-linguistic factors. The issue of creative word-formation is, however, especially relevant to the corpora that will serve as the core of this study, both of which are poetic in nature. For precisely this reason, to be able to concretely grasp the degree of productivity of a formation could provide indicators as to when productive or when creative word-formation is at work; see further under section 1.4.2.

#### 1.3.1.2 Productivity as a Statistical Notion

Granted that grammatical productivity is a matter of possibility or potentiality, the question then becomes how to "operationalize the notion of possibility" (Plag et al. 1999: 15). In effect, when attempting to determine how productive a given morphological construction is, one is trying to determine the likelihood that a speaker will select that particular morphological construction, as opposed to some other morphological or syntactic means of expressing the same concept. This must be what Bolozky (1999: 7) intends in writing that "lexical formation is first and foremost semantically based and concept driven" which seems to mean that speakers coin words in order to express a notion that a lexical item at hand cannot fulfill.[26] If the linguistic system offers the means to express the concept morphologically, then a coinage may come about; the relation to productivity concerns how likely it is that a given process will be invoked in order to express that concept.

Hence, if morphological productivity is the **likelihood** of employing a given construction, then productivity is understandable precisely as a probabilistic (statistical) notion, which makes Bolinger's definition above apt. While native speakers and linguists may have intuitions about the productivity of a construction, the genuine productivity of the construction, from an objective standpoint, is reflected in the extent to which formations are actualized.

At this point, I can offer a working definition of morphological productivity: it is *the probability that a given morphological construction will be used to express a given concept.* That probability is otherwise recognizable as the DEGREE OF PRODUCTIVITY.[27]

That this probability of usage depends on vast array of subtle factors, many of which may escape the notice of the linguist, seems self-evident. What can act as a first approximation

[26]The fact that a syntactic construction can often fulfill the same semantic and conceptual function as the product of a morphological construction, e.g., an agent noun *chair-builder* (with incorporated object) or a noun phrase with relative clause *a person who builds chairs*, is further evidence of the tight linkage between morphology and syntax. Morphology competes not only with morphology, but also with syntax. I believe that the fact that larger constructions are evidently susceptible to the measurement of productivity through the same means described in Chapter 2 is only further indicative of the interpermeability of the morphological and syntactic domains. The successful application of those same techniques to syntactic phenomena in Zeldes 2012 seems to prove the point.

[27]This degree of productivity could be equivalent to the weight of a markedness constraint that militates against the usage of some morphological process, such that higher-weighted processes are less likely to apply, all other things being equal.

to the DEGREE OF PRODUCTIVITY would then be the likelihood that a speaker will call upon a given morphological construction when the speaker has no access to a pre-existing means of expressing a given concept.[28] Data of this kind is readily available in the form of corpora, and therefore any language attested in the form of a sufficiently large corpus may be susceptible to the analysis of productivity. Thus, while all of the factors that directly account for productivity may remain hidden, at least directly grasping and utilizing productivity as a means of accounting for other linguistic phenomena should be possible.

For the moment, I will briefly consider some general factors that may often underlie and influence a construction's DEGREE OF PRODUCTIVITY.

### 1.3.2 Explaining Productivity

All attempts to account for productivity without a quantitative element look either to purely structural and categorical factors, or draw on analogy as the mechanism underlying productivity. That the terms 'morphological productivity' and 'analogy' can, at least in some instances, be applied to the description of the same phenomenon, is difficult to deny; however, to treat productivity and analogy as equivalent is helpful only insofar as one has a means to account for the (non-)application of given analogies. Since the means to directly explore and test the applicability of analogical domains do exist, I think that this connection can be fruitful, but I will delay the detailed exposition of those methods to Chapter 2. Structural factors likewise play a crucial role in establishing categorical boundaries for the application of a morphological process, and can indeed serve as factors alongside quantitative factors in a larger model of a language's morphology. But structural factors alone can neither directly capture nor gradiently relate the productivity of categories to one another, and have no hope of accounting for the extent to which productivity depends purely upon language use (i.e., that usage may beget usage, without any other interference). Inasmuch as usage or lack thereof impacts productivity, the way in which morphologically complex forms are processed is a crucial consideration; indeed, because this explanatory aspect of productivity is perhaps the most important, and has the clearest direct relation to the measurement of productivity, I will treat this matter separately in Chapter 3, making only a few summary remarks here.

#### 1.3.2.1 Structural Factors

In most discussions of morphological productivity that are oriented towards explaining productivity in terms of linguistic competence, syntagmatic restrictions on the applicability of processes play a major role. Perhaps the most common restriction encountered is the limitation of the type of base (i.e., syntactic category) to which a process can apply, e.g., the English prefix *re-* applies only to verbs, or the Italian suffix *-ità* ($\sim$ English *-ity*) applies only to adjectives. Since various structural descriptions of this sort are familiar, and described in

[28] Where a pre-existing means, i.e., a pre-existing lexeme, is available, it will usually act to block another semantically equivalent formation. Cf. fn. 4 above.

detail in other works on morphological productivity (cf. *inter alia* Bauer 2001: Ch. 5, Plag 1999: Ch. 3, Aronoff 1976: Ch. 3), I see no need to recapitulate the discussion here. Bauer identifies some plausible possible restrictions at all levels of the grammar, from phonology to pragmatics, in addition to the effects of blocking (e.g., the coinage of *furiousness* is dispreferred given the pre-existing and familiar *fury* that covers the same semantic ground). Plag's work, meanwhile, shows that restrictions on word formation are readily implemented as markedness constraints in Optimality Theoretic analyses.[29]

Less familiar are the sort of paradigmatic restrictions on productivity that van Marle (1985) has proposed. Van Marle's approach to productivity is fundamentally Structuralist in nature: he essentially claims that, for every concept that speaker needs to express, if more than one formation or construction is available, one formation will be the general case, while all other formations are special cases. The general case, in principle, can apply anywhere, granting certain syntagmatic restrictions, while the special cases apply to smaller parts of the domain that the general case covers. For example, the Dutch suffixes *-heid* and *-te* (= Eng. *-th*), which form abstract nouns from adjectival bases, would represent general and special cases, respectively. The instances where the general case applies should simply be all of the cases where the special case does not apply.[30] However, instances in which two or more competing forms actually occur (as is the case with many Greek aorists, for instance) refutes this notion, because, if the environment for the special case is present, then the application of the general case should be blocked. See further arguments on this point in Baayen 1989: 13–6.

Although, on the one hand, these sorts of structural constraints may help understand the behavior of morphological processes, they cannot fully account for the probability of a process' usage; I rather follow Baayen (1993: 183–90) in thinking that productivity is best grasped through observation of actualized forms, not restrictions on potential forms. A true description of productivity based on linguistic competence alone is literally impossible, since recursion in morphological derivation, just as in syntax, creates a theoretically infinite number of types. As Baayen (1989: 24–6) argues, trying to approach productivity from the perspective of linguistic performance requires consideration of the fact the extra-linguistic (socio-cultural) factors are at play in shaping the linguistic output that provides the data for the study of productivity.

[29] The problem of completeness in analysis (i.e., identifying all relevant factors) recurs here too, and all the more for deciding between very productive processes with few apparent restrictions.

[30] Conceptually, this approach is akin the dual-mechanism model of inflectional morphology (Prasada and Pinker 1993): wherever listed "irregular" processes do not apply, a general "regular" process steps in. The practical application of this idea is precisely refuted by the behavior of the Minimal Generalization Learner (see further 2.3). Take the case of English preterite formation: a rule of the form Ø → *-d* / __ is the general case, while other "irregular" preterites are special cases. However, while some forms may be generated by the true general case rule, many surface forms that look like the general case, may in fact, be "islands of reliability" (e.g., all verb stems ending in a voiceless fricative have the ending [-t]; cf. Albright and Hayes 2003), generated through a more specific rule, just like the special cases. In effect, to support a strict structural distinction between general and special cases is not empirically sound.

#### 1.3.2.2 Measuring Productivity

Although measuring morphological productivity is the explicit topic of Chapter 2, I introduce here two measures of morphological productivity that I believe to have theoretical and empirical validity. Both measures center around *hapax legomena*, forms that occur only once in some sample of a language (cf. further 2.2.2), as indicators of rare or novel forms that speakers encounter. These measures are corpus-based; they stem from work of Baayen (Baayen 1989, Baayen 1992, Baayen 1993). The core measure is $\mathscr{P}$, "productivity in the strict sense", which is calculated by the ratio of hapax legomena ($n_1$) belonging to some morphological category to the number of tokens ($N$) belonging to that category; this measure may be associated with the parsability of morphologically complex words. An additional measure is the ratio of a category's hapax legomena to all hapax legomena in the corpus, thus reflecting the proportion of potential neologisms and highly parsable forms that the category contributes to the language as a whole; Baayen labels this measure $\mathscr{P}*$. This brief description will suffice for the remainder of the present chapter.

#### 1.3.2.3 Productivity and Analogy

Already Saussure seems to have held that the production of both morphological forms and syntactic constructions reflected the same active linguistic process, which he put under the heading of ANALOGY (see de Saussure 1983: 221 ff.); as De Mauro (2003:451) remarks, "per S[aussure] sintagmi sono non sole le «parole», ma anche le «frasi», sicchè l'analogia è la fonte della creatività della lingua, la via attraverso cui la lingua genera l'insieme teoricamente infinitivo delle frasi."[31] Saussure further observes that "changes" in the surface form of words not attributable to sound change "are the same as what we call 'creations'" (de Saussure 1983: 226). If Saussure is correct on this point, then indeed, the results of alterations to existing forms (e.g., Lat. *sororis* : *soror* :: *honoris* : *honos* >> *honor*) come out of the same psychological mechanism that generates neologisms, forms that in replace nothing (*ibid.*: 225). The basic notion underlying Saussure's view is that both alteration and creation are possible through pattern extension.[32] The chapter on "Analogic Change" in Bloomfield (1931 [1984]) also groups together replacement and productive word-formation.

Similarly, other scholars of historical linguistics follow Saussure, and recognize analogy and productivity as fundamentally the same. Hock (1991: 173 ff.), for instance, points out that productive creations very clearly fall out from normal four-part analogies, and like Saussure, acknowledges that such analogies can successfully model both the reshaping of old forms and the generation of entirely new forms. Moreover, some definitions of analogy cannot escape the use of the term 'productive'; Fortson (2010: 6) states:

> The replacement of the old plural *kine* by *cows*, and of the old past tense *holp* by *helped*, are examples of a lexical change; in cases such as these an old irregular

[31] "for S[aussure], syntagms are not only words, but also phrases, insofar as analogy is the source of creativity of a language, the means by which a language generates the theoretically infinite set of sentences."

[32] "An analogical form is a form made in the image of one or more other forms according to a fixed rule" (de Saussure 1983: 221).

> form (containing morphemes or morphological processes that are **no longer productive** [emphasis mine]) is replaced by a regular form, by a process called *analogy*.[33]

The conclusion that analogy and productivity are merely two sides of the same coin thus looks attractive; Becker 1993 is a thorough recent defense of this claim on the basis of German word-formation patterns. One major point on which productivity and analogy would seem to differ sharply, however, is that the productivity of a word-formation rule readily admits of scalar and statistical interpretations; analogical replacements, on the other hand, often give the impression of being precisely random and unpredictable.

Therefore, scholars working on morphological productivity are not universally comfortable with equating of analogy and productivity. In the main, the concern is that word-formation can be readily captured with rule-like descriptions, whereas analogy evades description in terms of rules. For a work like Becker 1990, if morphological rules and analogies reflect the same psychological mechanism, then analogy and productivity are perforce the same as well. Plag (1999: 17) worries "this has the considerable disadvantage that it is left unexplained why some analogies are never made, but others are frequently observed." Bauer (2001: 97–8) is willing to accept the possibility of unifying productivity and analogy, but feels that the entire matter turns on whether morphology in general operates with independent rule-based and analogical mechanisms, and hence is not to be decided on the basis of word-formation processes alone.

The seeming unpredictability of analogy, however, is now all but a relic. Several plausible competing models, with computational and statistical implementations, which make precise, well-defined predictions, are available: most prominent are Analogical Modeling (AM; Skousen 1989, Skousen et al. 2002), Memory-Based Learning (MBL; Daelemans and van den Bosch 2005, Keuleers 2008), and Minimal Generalization Learning (MGL; Albright 2002b, Albright and Hayes 2003). I will review these different approaches, and offer reasons to prefer MGL under 2.3. Chapter 5 will ultimately show that a model of analogical rules like MGL makes predictions similar to corpus-based measures with respect to derivational morphology, while Chapter 8 shows such models to be preferable, perhaps indispensable, for the meticulous treatment of inflectional productivity.

#### 1.3.2.4 Language Processing

The principle that high token frequency word forms are resistant to analogical reshaping or lexical replacement is a stable piece of wisdom in historical linguistic work. At the same time, a large body of psycholinguistic work exists, which has demonstrated that speakers are indeed sensitive to word frequency effects; in particular, word forms with high token frequencies have significantly faster response times in lexical decision tasks.[34] See Baayen

[33] Note that Fortson leaves the term "productive" itself totally undefined, but assumes that its lack is a precondition to analogical replacement.

[34] Lexical decision tasks present a subject with a word stimulus (visual in most experiments, but occasionally aural), and ask the subject to determine whether the form is an existing word of the subject's language. Some

1989: Ch. 7 for a good summary of the relevant literature up to that date. The historical and psycholinguistic evidence combined makes clear that even words that are etymologically morphologically complex are subject to full-form storage (i.e., listing in the "dictionary"), and may be produced and processed through lexical access, rather than morphological (de)composition, depending upon frequency. In particular, Frauenfelder and Schreuder (1992) have proposed a "morphological race model", under which access from memory and morphological parsing compete to provide a representation for an incoming word.

The significant implication for the study of morphological productivity here is that morphologically complex words that rely on memory for lexical retrieval and production, because of their high token frequency, may not contribute to the psychological representation of the morphological construction that the word (etymologically) instantiates. Hence, if a type of formation primarily consists of words with high token frequency, the construction itself may have little independent representation in memory, be it a word-formation rule or other psychological entity. Consequently, that word-formation rule is less likely to be called upon for the production of neologisms. Ironically, then, the more successful that an individual word form is (i.e., it comes to have great pragmatic utility in the language), the less that the word contributes to the productivity of the morphological process that it instantiates. This situation also explains the importance of *hapax legomena* in the practical measurement of morphological productivity.

The general point is that some demonstrable relationships between lexical access and word formation can and should be a consideration in an understanding of morphological productivity, and may play a crucial role in the historical trajectories of individual word forms. More detailed evidence and specific implications will be treated in Chapter 3.

## 1.4 The Application of Productivity in Historical Linguistics

Given that the means to describe, measure, and account for the productivity of morphological categories are available, the question then arises: what benefit do such procedures potentially bring to the historical linguist? Insofar as the description of productive categories makes up part of the grammatical description of a language, to concretely assess the productivity of various formations in a language is of inherent interest to linguists concerned with that language. For Indo-Europeanists, for instance, an account of productivity in the derivational morphology of Ancient Greek and Vedic Sanskrit is valuable precisely because it contributes to the grammatical description of those languages. Similarly, the documentation of changes in the productivity of a morphological process is as much a part of a language's history as the documentation of sound changes. In Chapter 5, for example, we will see quantitative evidence that, during the 1st millennium BCE, sigmatic aorists in Greek continued to gradually grow in degree of productivity, while supplanting the other aorist categories, at least on a type-by-type basis.

---

evidence for morphological productivity emerges from studies that have shown that speakers are more likely to interpret words with productive morphology, though not recorded in standard dictionaries, as being actual words of the language.

However, I wish to take a further step forward, and to propose specifically how the information gleaned via productivity measures can aid in the resolution of historical linguistic problems. Specifically, the problems that confronted the editors of the LIV and NIL, as seen in 1.2.2, could concretely benefit from the results of productivity measures. Below I make some suggestions as to how probabilistic measures of productivity might be interpreted as the probability whether a given word is inherited (from some older phase of the language) or is a recent creation. The assumptions that the probabilistic measurement of productivity entail can also help to develop a methodology for the interpretation of individual words based on their token frequencies and the frequency of the type to which the word belongs.

Finally, I will make an inquiry into the potential utility of productivity measures for comparative reconstruction. Given measures of productivity for two morphologically related formations in two related languages, perhaps some interpretation as to the productivity of that formation in the most recent common ancestor of those languages is possible. If the reconstruction of degrees of productivity is feasible, then the study of the linguistic prehistory of languages without attested ancestors may become possible along a further dimension. I also consider the question of whether frequency information be used to make predictions about the prehistory of specific forms.

### 1.4.1 Existing Diachronic and Historical Studies of Productivity

The diachronic study of morphological productivity is open, though not altogether untrodden, territory. The study of productivity in historical corpora, as opposed to corpora built on linguistic data from the 20$^{th}$ and 21$^{st}$ centuries, is hardly common, and seems so far to be essentially limited to study of the history of English. These existing studies do already illustrate, however, that the meaningful analysis of productivity, using solely corpus-based data, is possible, and from that point of view, they encourage further studies.

A preliminary problem concerns whether a given corpus properly conveys synchronic or diachronic information about morphological productivity in a language. Cowie and Dalton-Puffer (2002: 421), in addressing this point, conclude that no strict answer is possible: "what counts as synchronic and a time-point and what as diachronic and a time stretch is ultimately a matter of definition and methodological necessities." In principle, the same body of data could furnish both synchronic data on productivity, taking the corpus as a whole for the time-period that it covers, and diachronic data on changes in productivity within that time-period, by dividing the corpus into sub-corpora.

While Plag (1999: 101) may be justified in holding that his corpus of English for the years 1900–1985 is "small enough to exclude major diachronic developments", one could certainly find smaller changes in the degree of productivity of some formations. Indeed, Baayen and Renouf (1996: 80–2) note an apparent increase in the degree of productivity of English *-ness*, *-ly*, and *un-* over just a four-year period (1989–1992); however, only because the corpus that they employ is so large (80 million words) is that sort of fine-grained study possible, without being able to attribute possible differences to sampling error.

Baayen and Renouf 1996: 80–2 also merits discussion as an example for the sort of distri-

butions that indicate changes in the productivity of a formation. The authors observe that the number of new hapax legomena with respect to the number of tokens sampled over time increases for the English suffixes *-ness* and *-ly*, and the prefix *un-*. The degree of productivity value for those formations (as measured by Baayen 1989's $\mathscr{P}$) is grows gradually larger over the 1989–1992 timespan of the sample. In contrast, the number of new hapaxes for the prefix *in-* and the suffix *-ity* remains roughly constant as more tokens accumulate over time, thus indicating that these formations maintain the same degree of productivity. Although this method seems ideal for the diachronic tracking of productivity, its easy and confident implementation is not possible where data are too sparse, or where a definite chronological arrangement of texts is unknown.

I can briefly report here on other studies of productivity with a diachronic bent:

- Bauer (2001: 163–72) undertakes a brief comparative study on the productivity of a suffix **-dōm* common to many Germanic languages (e.g., English *wisdom*, German *Weistum*). Bauer does not attempt to measure the productivity of **-dōm* across the the languages that he samples (English, German, Dutch, and Danish), but rather tries to chart different domains of productivity for the affix in the different languages. Specifically, Bauer claims that "restrictions on bases can change as part of the diachronic process of language change," based the observation that new forms in Germ. *-tum* and Eng. *-dom* are limited to nominal bases, though forms with verbal and adjectival bases exist in all four languages. Although interesting for the history of this particular affix, Bauer's examination does not clearly demonstrate how to study morphological productivity diachronically.

- Two other studies on English material, Cannon 1988 and Aronoff and Anshen 1998, both rely on type counts of different formations taken from dictionaries. Using dates of first attestation in the OED, the authors count the number of words listed with the suffixes *-ness* and *-ity* by century of attestation in order to discuss the relative productivity of those suffixes over time. While the coverage of the number of types for earlier periods of English in dictionaries is perhaps fairly complete, coverage for later periods is certainly only partial. Both Baayen and Renouf (1996: 70) and Cowie and Dalton-Puffer (2002: 423–4) emphasize the inadequacy of dictionary-based studies, so Cannon and Aronoff & Anshen's studies do not make for good models.

- Dalton-Puffer (1996) attempts to track the productivity of several English nominalizing affixes in the period 1150–1420. Dalton-Puffer focuses on the respective type and token counts of the affixes in three sub-corpora (1150–1250, 1250–1350, 1350–1420) as a means of tracking changes in productivity. The large increase in the number of types of both the suffix *-ation* between the 1250-1350 and 1350-1420 periods (20 and 138 types, respectively – though the 1350-1420 corpus has approximately twice the number of tokens as the 1250-1350 corpus) suggests an increase in the productivity of that suffix. However, recall the conclusion of Baayen (1989: 41) that concerning type frequency alone as a measure of productivity: "it is impossible to extract information concerning the number of possible items of a morphological class *S* from the observed num-

ber of types *V*." Dalton-Puffer may, then, have discovered an increase in the pragmatic usefulness (Baayen 1989's $\mathscr{U}$; see further 2.2.2) over time, but whether that entails a similar increase in strict productivity is unknown.

- Two recent studies of Early Modern English (Březina 2005 and Säily 2008) have made effective synchronic use of Baayen's $\mathscr{P}$ and $\mathscr{P}*$ measures to substantiate claims about differences in gender and/or register of different word-formational processes (*in-* and *un-* and *-ness* and *-ity*, respectively). Both of these studies rely on Nevalainen et al. 1998 as their corpus, though different portions thereof; Březina works from only 450,000 tokens, while Säily works from 1.4 million tokens. Březina's study is interesting for present purposes in that it seems to obtain plausible results using a corpus even smaller than the 600,000 token Eindhoven Corpus used in Baayen 1989 and Baayen 1992. Säily's work also directly investigates type and hapax accumulation curves (see Säily 2008: 66–70 for the exact procedure), as a non-parametric representation of the growth curve for a morphological formation. These curves allow Säily to demonstrate whether statistically significant differences in the productivity of *-ness* and *-ity* exist between subcorpora (divided on the basis of class, gender, etc. of the writer). Both type and hapax accumulation curves model the growth curve in the way that Baayen's statistic $\mathscr{P}$ is supposed to capture; similar vocabulary growth curves serve as a device for the comparison of productivity cross-categorically in Chapters 5 and 6.

Thus, while a not insignificant amount of work on productivity in historical corpora and diachronic changes in productivity has appeared, the usage of productivity measures for the purposes of establishing the history of morphological formations in the unattested prehistory of a language would be new ground. These existing studies, however, should give confidence that the methods to be described in Chapter 2 can contribute to such study.

### 1.4.2 Archaisms, Neubildungen, and Nonce-Formations

In describing the history of a language's derivational morphology, the historical linguist is interested in separating which forms and types of forms must be old (inherited from some older stage of the language), and which must be new (generated synchronically). Precise awareness of productivity at a synchronic phase (or more exactly, the largely coherent grammar that a corpus presents) immediately provides a useful prospective on this problem. Namely, the various frequency statistics can not only help to determine whether a kind of formation is productive or not, but the extent to which it is productive (or not), and thus the likelihood that a specific form has been generated productively.

On the other hand, the fact that a form belongs to a productive morphological category does not strictly exclude the possibility that it is old, since productivity not only generates new forms, but preserves existing forms. Consequently, a lack of other comparable forms in related languages, or perhaps the specific derivational pattern of the form, are necessary to confirm a form's recent character.

To anticipate some results from the case study in Chapter 5, we can examine a few of cases from aorist formation in Homeric Greek. The $\mathscr{P}$ values for the root, thematic, and

sigmatic aorists are .000738, .00332, and .0284, respectively. For the moment, I take the threshold for even a marginal degree of productivity in the Homeric corpus to be $\mathscr{P} = .001$ (i.e., one hapax legomenon per 1000 types) in a corpus of that size.[35] Hence, the possibility that a root aorist reflects a relatively recent formation is almost entirely to be excluded. This fact can contribute to the reconstruction of several aorist forms.

The LIV reconstructs an IE root **gem-* 'press (together), grasp' most confidently on the basis of Grk. γέμω [gémɔː] 'be full', Lat. *gemō* 'sigh', and OCS *žъmǫ* 'press'. Semantically, the 3.sg.mid. root aorist γέντο [génto] 'grasped' (5× *Il.*) evidently belongs with the Slavic, which also has 3.sg. *žê* that conceivably continues a root aorist. The LIV, however, marks the reconstruction of a root aorist as uncertain.[36] The degree of productivity associated with root aorists alone makes the stem γεν- a likely archaism.[37] I would therefore reconstruct a root aorist stem **gem-* to the last common ancestor of Greek and Slavic with confidence.[38] Indeed, the judgment of (Snell 1979-2010: s.v. γέντο) concurs with my judgment: γέντο [génto] is a "Reliktwort".

Slightly trickier is the case of the root aorist stem λεκ- in 3.sg.mid. λέκτο [lékto], 1.sg.mid. ἐλέγμην [elégmɛːn], which occur alongside a sigmatic aorist stem λεξα- [leksa-]. Here, the LIV is simply in error in following the claim of Harðarson (1993: 205), and the data on the respective productivity of root and sigmatic aorists in Homer is ignored at peril. Harðarson cannot argue on evidence internal to Homer that the stem λεξα- [leksa-] is indeed older, and that λεκ- [lek-] is an innovation; he instead depends on the testimony of the Lat. perf. *lēgī* and (*intel-*)*lēxī* as evidence for an old sigmatic aorist. However, Jasanoff (2012) has explained *lēgī*, by comparison to Alb. *mblodhi* and Toch. A *lyāka*, as continuing an asigmatic Indo-European preterite formation with *[ē]; the Lat., Alb., and Toch. forms thus continue an IE *[lēĝ-]. The virtual impossibility that the root aorist stem λεκ- [lek-] is not old in turn has the benefit of bolstering Jasanoff's proposal, since his "long-vowel preterites" might in origin reflect IE imperfects; Greek and Latin together would thus offer evidence for IE pattern pres./impf. *[lēĝ-] alongside aor. *[leĝ-].[39]

Concrete data concerning productivity might also be brought to bear on the analysis of so-called "nonce formations" or "Augenblicksbildungen". The term is especially common in the treatment of Vedic forms that appear to violate standard rules of Vedic word-formation;

[35] Cf. section 5.3.

[36] The full grade in γέντο may be unexpected, but if separated from the present γέμω [gémɔː], which is activum tantum, then the aorist γεν- would be medium tantum, like κεῖτο [keîto] 'lies', and a full grade middle would be less worrisome, from an IE perspective. Cf. also Melchert 2014 for arguments in favor of regarding full-grade middles as relatively normal formations.

[37] Furthermore, γέντο is the only form attested to that stem, and is very much restricted metrically: it falls only ever after the bucolic caesura. This metrical position is, however, the most common position of words with the metrical shape —⏑ (cf. O'Neill 1942: 140).

[38] Given the semantic divergence, whether the presents Grk. γέμω [gemɔː] 'be full' and Lat. *gemō* 'sigh' should remain with this aorist is unsure. Vine (2007) argues further that γέμω and *gemō* should be kept apart, and that *gemō* instead belongs with the Gk. perf. γέγωνε [gégɔːne] 'cry out'.

[39] See, however, Chapter 8 and Sandell 2014b, on the possibility that such rightly reconstructible surface long vowels might reflect a phonological pattern of reduplicated forms, i.e., *[lēĝ-] would result from underlying /Red-leĝ-/.

see Knobl 2009: 21–43 for a recent discussion, and an attempt to classify some Vedic "nonces" into categories. In principle, a "nonce" could have one of several grounds:

- it is an archaism, reflecting a dead word-formation rule, of which nearly all tokens have disappeared.
- it is a innovation, reflecting a very new word-formation rule, of possibly high productivity, all of whose types have few tokens, and which itself has few types.
- it is the deliberate creative coinage of a speaker (see 1.3.1) above.

If a creative coinage, then the analysis of the form falls outside the domain of linguistics proper. The other two possibilities must, however, be eliminated – genuine nonces ought to be not only unique types, but unique tokens as well. A recurrent form might suggest the first possibility, an archaism, while an innovative category (even if fleeting itself) must attest more than one type. Perhaps the productivity of the morphemes themselves involved in nonces might help account for the existence of some "nonces" as extensions of the environment for the application of a word-formation rule or schema, in either erroneous or deliberately creative ways.

### 1.4.3 Productivity in Reconstruction

The principal task of historical linguistics is simply to describe and explain the history of a language. Reconstruction is a necessary tool towards that end just in case no older phases of a language are attested in the historical record. The application of the Comparative Method, so successful in phonological reconstruction, is often less straightforward with respect to morphology. Just in case a morpheme formally and functionally agrees across languages, no real difficulty arises; for instance the reconstruction of a nom.sg.anim. morpheme */-s/ for PIE is totally unproblematic. On the other hand, the reconstruction of the PIE 1.sg.pres.act. ending to present stems formed with thematic suffix */-o/e-/ is slightly trickier: Lat. *-ō* (*ferō*), Grk. -ω [-ɔː] (φέρω [pʰerɔː]), and Av. *-ā* (*barā*) agree in the reconstruction of IE *[-ō] (= */-o-$h_2$/), whereas Ved. has *-ā-mi* (*bhárāmi*). The particular solution in this case is not difficult to grasp – *-mi* in Vedic is somehow imported from the inflection of athematic verbs that have *-mi*.[40] The point is merely that historical morphology is a more of a moving target than historical phonology, and therefore, the morphological composition of a word cannot be taken at face value; the usage of Transponat reconstructions demonstrates awareness of this point.

At the same time, to accurately describe and explain the history of a form or of a morphological category requires some starting point, which is obscure when that starting point lies in a reconstructed phase of a language. For example, although adjectives formed with a

[40] Interestingly, to make the analogy transparent, derivation of the present stem must be clearly separated from the inflectional endings; taking the 1.pl.pres. as a base, a mapping *-masi* → *-mi* / X_ ] overtakes the mapping *-masi* → Ø/ ā_] – that is, the older 1.sg. *-ā* is taken as a stem formant *-ā-*. See Hill 2012 for an altogether different view.

suffix */-ró-/ certainly existed in Indo-European, since they are known in Greek, Vedic, Latin, etc., how diffuse this suffix was in Indo-European, and exactly how its specific distribution in the particular Indo-European daughter languages came about, is less clear. In order to discuss the place of some morphological formation in the proto-language, one must proceed from the lexicon of the proto-language, just as one would with an attested language; however, because the lexicon of the daughter languages is subject to lexical renewal and replacement, a clear picture of that necessary lexicon of the proto-language is itself difficult to obtain.

It is precisely on this point that concrete measurement of productivity, and the systematic discovery of the domains of productivity for a kind formation, can be of aid. Consider the following scenario:

> Stem A in Language 1 consists of a root (R1) and a suffix (S1). Suffix S1 is a productive suffix of the language. Part of the distribution of suffix S1 shows that it often replaces the unproductive suffix S2. Stem A fits the domain for the productivity of S1. We can then posit a stem *A, consisting of R1 and S2. Stem *A may be projected back to a proto-language, depending upon the existence of stems matching *A, and the respective productivity of S2 in sister Languages 2, 3, etc.

In effect, assessing and measuring morphological productivity internal to a language may offer the means to "peel back" the derivational history of an actually attested lexical item. This procedure gives insight into the lexicon of the proto-language, and thus the possibility of tracing the history of morphological categories in greater detail. Simple comparison already permits us to identify likely reconstructible forms; incorporating frequency data and measures of productivity for particular categories, it may be possible to make predictions about the reconstructibility of a given form where the normally necessary comparanda may be lacking.

# CHAPTER 2

# Measuring Morphological Productivity

## 2.1 Chapter Goals

In Chapter 1, I proposed an explicitly probabilistic definition of morphological productivity: *the probability that a given morphological construction will be used to express a given concept*. While the concepts that persons wish or need to express are external to grammar (though see below on PRAGMATIC USEFULNESS), the grammar of a speaker delimits the means that the speaker adopts in order to express those concepts. At the same time, factors perhaps external to the grammar proper will inevitability affect the frequencies of specific words and morphological constructions, and that primary linguistic data will in turn shape speakers' grammars.

I used section 1.1 to point out some of the difficulties in works on Indo-European historical linguistics that result from the lack of a well-defined measure of "productivity". In particular, the absence of measurement both makes difficult the evaluation of claims regarding productivity itself, and also renders conclusions drawn from those claims more uncertain. Furthermore, if I am correct that the type frequency of a formation is the typical basis for intuitions concerning productivity in corpus languages, then some of those intuitions rest on unstable ground, because type frequency alone does not consider the *rate* at which new types are introduced into the language.

The objective of this chapter is to present methodologies that can serve both the definition of productivity that I offer, and be of use in the practice of historical linguistics, especially in the case of corpus languages. In the main, I propose to measure morphological productivity through the methods that R. H. Baayen, going back to his dissertation (Baayen 1989) and an early paper derivative thereof (Baayen and Lieber 1991), has developed. Baayen's measures are grounded in the statistical analysis of linguistic corpora, and thus simultaneously escape the quandary that factors beyond the linguistic system itself pose for the study of productivity by incorporating them, and hence find suitable application in languages that exist only in the form of corpora.[1] Baayen's measures of productivity are the topic of 2.2. In order to explain how and why Baayen's measures can credibly function as measures of morphological productivity, it will be necessary to treat the basic properties of word frequency distributions in language corpora.

While measures like Baayen's $\mathscr{P}$ ("productivity in the strict sense"; the ratio of items oc-

[1] Other major works on the problem of morphological productivity (Plag 1999, Bolozky 1999, Bauer 2001) have discussed and employed the work of Baayen from the early 1990's, and some have already considered and applied Baayen's measures in historical studies (Cowie and Dalton-Puffer 2002, Säily 2008; cf. 1.4.1 above).

curring just once to all tokens in the sample of some category) provide a convenient means of comparing the productivity of morphological categories within a corpus (and with appropriate transformations, possibly between corpora), they are relatively gross, and by nature, global in scope. Conversely, to examine the effects and potential impacts of productive processes on individual lexical items is possible with analogical learning models; I discuss some potential candidates at 2.3 and select the MINIMAL GENERALIZATION LEARNER (MGL) as the best available tool to this end. The MGL, despite some limitations, is particularly well-suited to the study of analogies that are principally formal (i.e., morphophonological) in motivation. It is not my objective in this chapter to establish mathematical proof or to account for all technical workings of the methods discussed in 2.2 and 2.3, but rather to provide a minimum of formalization, intuitive illustration, and motivation for their employ in the present work. References to more details of underlying theory will be given throughout.

## 2.2 Productivity and Word Frequency: Statistical Bases of Word Frequency Distributions in Corpora

The foremost object of this section is to summarize the development of several corpus-based measures of morphological productivity. The basic measure $\mathscr{P}$ is discussed thoroughly in Baayen 1992, while Baayen 1993 introduces introduces a further measure $\mathscr{P}*$, and makes initial suggestions as to the relation between those two measures of productivity and a psychological measure of the "activation level" of a morphological formation, $\mathscr{A}$.[2] These measures undergird, for instance, the study in Baayen and Renouf 1996, which documents small changes in the productivity of some English derivational morphemes.[3]

With a practical understanding of the core productivity measures to be employed in the case studies of Part II in hand, I then review the essential facts concerning frequency distributions in corpora. I include this information in order to offer some further background into the theoretical bases that underlie the quantitative study of productivity. In particular, I hope to make clear the underlying statistical properties typical of word frequencies. This material, covered in section 2.2.3, can be passed over, though is essential to the discussion of how "pragmatic potentiality", $\mathscr{I}$, is calculated, at 2.2.3.1.

The already expansive literature on the corpus-based study of morphological productivity that makes use of the measures $\mathscr{P}$, $\mathscr{P}*$, and $\mathscr{I}$ helps us to understand what these

[2] The psycholinguistic validation of these measures will be treated in Chapter 3.

[3] Although numerous scholars writing on morphology (for a short survey, see Baayen 1992: 111) have made proposals concerning the relation of productivity and word token and/or type frequency, to my knowledge, Aronoff is the only other scholar who has proposed a measure of productivity that has a hope of withstanding theoretical scrutiny. Aronoff's (1976: 36) idea is that "we could arrive at a simple index of productivity for every WFR: the ratio of possible to actually listed words." Aronoff's proposal here, recalls Saussure: "Any word that I improvise, like *in-décor-able* already exists potentially in the language…Its actualization in speech is an insignificant fact in comparison with the possibility of forming it" (de Saussure 1983: 227). Aronoff's proposal is functionally equivalent to Baayen's measure $\mathscr{I}$ treated below (cf. discussion in Zeldes 2012: 86–90). Determining an estimate of the theoretical number of types belonging to a morphological category poses a special, but not irresolvable, problem, which is treated in 2.2.3.1 below.

measures mean, how they can be practically employed, and what assumptions or transformations of the underlying data must be undertaken in order to compare them. In particular, I will discuss some proposals made by Gaeta and Ricca (2006), Gaeta (2007), and Zeldes (2012) on the interpretation of these measures at 2.2.3.2, in order to refine my conception of how these measures of productivity are most appropriately deployed.

### 2.2.1 *hapax legomena*

Before turning to discussion of the workings of some productivity measures, I must first describe in some detail the theoretical role of the *hapax legomenon* ("read once"; pl. *legomena*), which I label as $n_1$.[4] *hapax legomena* here will refer specifically to word types that occur only once in some sample of words.

In philological and linguistic work on old Indo-European languages, the term 'hapax legomenon' often connotes an archaism; the fact that a word occurs uniquely (in all records of the language, or in a particular text) is often taken as an indicator of an old word that has nearly fallen out of use altogether.[5] Conversely, hapax legomena may be neologisms, generated through productive word-formation processes, or less often, creatively formed words (in the sense of 1.3.1.1 above). Intuitively, hapax legomena are crucial for the measurement of morphological productivity, because they can be reflective of neologisms, and neologisms are perforce indicative of productivity: a speaker will only generate a neologism based on a pattern that is productive for him. Baayen (1993: 189) states that "as sample size increases, the proportion of neologisms among the hapaxes will increase. Hence, the probability of encountering neologisms is measured indirectly by means of the probability of encountering hapaxes."

The hapax legomena that appear in a given corpus are not properly to be understood as neologisms in a language in the strict sense (though a particular hapax legomenon might happen to be such a neologism), but merely representative of the kinds of neologisms that should occur in a yet larger sample of the language. The reasoning follows that, if neologisms are indicative of productive morphological processes, then a measure of the rate at which neologisms belonging to a given type of morphological formation enter the language should constitute a measure of that morphological process' productivity. Thus, in the context of work on morphological productivity, a hapax legomenon is defined strictly with respect to a given corpus, which is a sample body of a language constituted by $N$ tokens drawn from the entire population that is that language.[6] Any form that occurs exactly and only one time, just in that particular corpus, is a hapax legomenon with respect to that corpus, regard-

[4] In some other works, and in some figures in this work, the alternative label $V_1$ appears.

[5] Normally, if a word happens to occur in a given text only once, but is otherwise known with regularity in the language, then the word is not to be understood as an archaism. If the corpus in which such a word is a hapax legomenon is an older document, then this situation may reflect a word that has become more popular at a later time.

[6] From the point of view of probability theory, we must consider a corpus as a sample drawn from a population that is representative of that population. However, in practical terms, what the population itself would be is difficult to define.

less of whether the form may regularly occur outside that corpus. Hence, I will use 'hapax legomenon/a' throughout this study to refer to a form or forms that occur(s) precisely one time in a given corpus; whether the form is attested elsewhere or not is irrelevant.

Indeed, the hapax legomena of a corpus, from a psychological point of view, may reflect more than an approximation of the genuinely novel lexical items created in that language. Consider that each speaker of a language, in the course of his lifetime, hears and produces some finite number of utterances, and among those utterances, some forms will have occurred but once. Many words that have been catalogued by lexicographers, thus implying that they have obtained some degree of purchase in the language, may in fact happen to be experienced by a speaker only once in his lifetime. Thus, within the corpus of linguistic experience for that speaker, many of his hapax legomena might be considered perfectly "regular" words of the language. Analogously, a given corpus may contain various hapax legomena that are otherwise reasonably well-attested in the language.

### 2.2.2 Measures of Productivity: From "Strict" $\mathscr{P}$ to "Potential" $\mathscr{I}$

We now turn to some concrete statistical measures of morphological productivity, which can be implemented with data from any sample corpus. Baayen (1989: 25–6; 57–68) proposed three different measures for distinguishing three different facets of morphological productivity using frequency distributions in language corpora, which I describe in turn here.

1. $\mathscr{P}$: productivity *stricto sensu*, "the readiness with which a rule is put to use," is measured as the ratio of hapax legomena ($n_1$) belonging to a morphological formation to the number of tokens (the category-conditioned $N$) belonging to a that formation.[7] This ratio expresses the probability that a new token (i.e., one more token added to the corpus) belonging to a particular formation will instantiate a new type belong to that formation. In terms of conditional probability: $\frac{n_1}{N} = P(x \text{ is new} \mid x \text{ is in category } X)$. $\mathscr{P}$ is the probability that a token $x$ is new given that $x$ belongs to category $X$, where the probability that $x$ belongs to category $X$ is given by the ratio of tokens belonging to some category (the category-conditioned $N$) to the number of tokens in the entire sample (the corpus $N$).[8]

   For example, let us say that $5000$ tokens have been sampled in a corpus of Ancient Greek, and that $100$ tokens belonging to the category of SIGMATIC AORIST have occurred, with $20$ different types, five of which are hapax legomena; when the $5001^{\text{st}}$ token is sampled, given that that token is a sigmatic aorist (which is $100/5000 = 0.02$), the probability that it will be a new type (i.e., not one of the $20$ already known types)

[7] We may also refer to $\mathscr{P}$ as the "category conditioned degree of productivity" (Hay and Baayen (2002: 18)), since it measures productivity with respect to one specific morphological category at a time.

[8] To be clear, $N$ as used throughout this work may have two readings: either the total number of tokens belonging to a corpus, or the total number of tokens belonging to some linguistically defined category within a corpus. Which sense is intended should always be reasonably clear; in general, $N$ may be read simply as 'token frequency'.

is $0.05$ $(5/100)$. This probability describes the *rate* at which types are accumulating in a category.

That rate of type accumulation immediately makes sense of the relationship between type frequency and productivity: the sheer number of types belonging to a category is not as crucial as how often new types are encountered. Based on a corpus, one can plot the ratio of types $(V)$ to tokens $(N)$ as ever more tokens are sampled; this graphical representation is a *vocabulary growth curve* (VGC) that illustrates the successive change in rate of growth for a category. Baayen defines $\mathscr{P}$ as ($\Delta$ is "change in")

$$\mathscr{P} = n_1/N = \Delta V/\Delta N \tag{2.1}$$

where $\Delta V/\Delta N$ gives the slope of the growth curve for the number of types belonging to a morphological category at $N$. More explicitly, the rate of vocabulary growth for a given category is the derivative taken to the ratio of types to tokens at some token sample; $\mathscr{P}$ is normally expressed as the rate of change in $V/N$ for the maximum $N$ of a category. Figure 2.5 below gives a sample VGC for the Dutch suffix *-heid* (functionally similar to English *-ness*) in the Eindhoven Corpus (EC). Where $N = 1000$, $V = 299$, the slope of the tangent to the VGC is $0.177$, that is, $\mathscr{P}$ would equal $0.177$ if $1000$ were the maximum number of tokens of *-heid* found in the EC. In actual practice, $\mathscr{P}$ for the suffix *-heid*, the rate of increase in types with the suffix *-heid*, will be calculated as the number of hapax legomena with the suffix *-heid* at the maximum $N$ of $2251$.

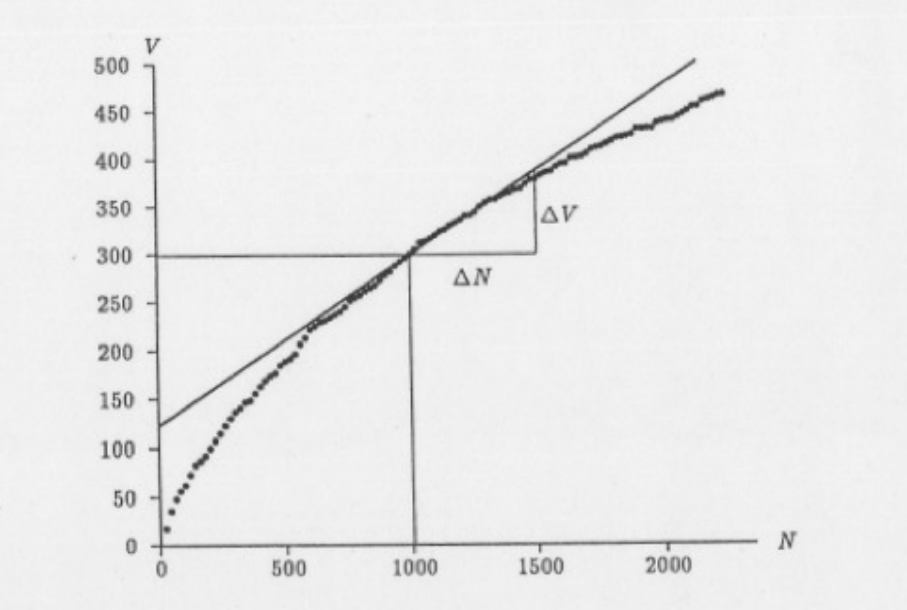



Figure 2.1: From Baayen 1992: 113: "The growth curve of *-heid* in the EC ($N = 2251$, $V = 446$). The growth rate for sample size $1000$ can be expressed in terms of the slope $\Delta V/\Delta N = 0.177$ of the tangent to the curve in the point (1000, 299)."

The VGC plotted for *-heid* in the EC is an *empirical VGC*, because it plots each new type as one scans through the tokens of *-heid* in the order in which they occur in the EC. The VGC is appears relatively smooth because new types of *-heid* appear at sufficiently regular intervals. In contrast, consider the empirical VGC for Ancient Greek nouns with the suffix [-si]-/[-ti-] (e.g., ῥῆσις [rɛ̂ːsis] 'discourse', μάντις [mántis] 'prophet') that occur in Homer at Figure 2.2.[9] The curve in this case instead appears much more jagged, as progressively longer plateaus in growth appear as new tokens are sampled.

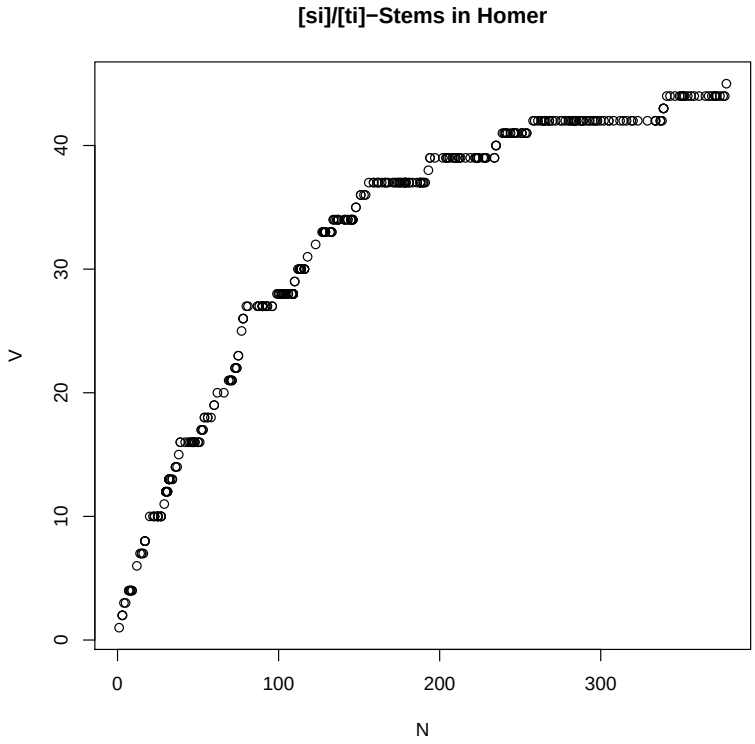



Figure 2.2: Empirical Vocabulary Growth Curve for [si-]/[ti]-Stems in Homer

The curve appears jagged simply because word frequencies are discrete units: the occurrence of a new type is a sporadic event, which interrupts the more gradual accumulation of tokens. A smoother continuous representation of the growth curve, however, can be obtained by the method of binomial interpolation (cf. Baayen 2001: 64–9, Evert and Baroni 2008), as shown in Figure 2.3, which represents the rate of vocabulary

[9] See Chapter 7.4.1 for more details on this category. This growth curve was created as follows. First, I prepared an XML-structured file of the *Iliad* and *Odyssey*, where all word tokens were given a tag for morphological category; I tagged by hand all tokens of [si-]/[ti]-stems. A separate R script that I wrote was then used to parse the XML document token-by-token, and create a data frame with accumulating token ($N$), type ($V$) and hapax legomena ($n_1$) for the category. The plot here then derives from the paired token and type frequencies in that data frame.

growth as continuously distributed across all tokens.[10]

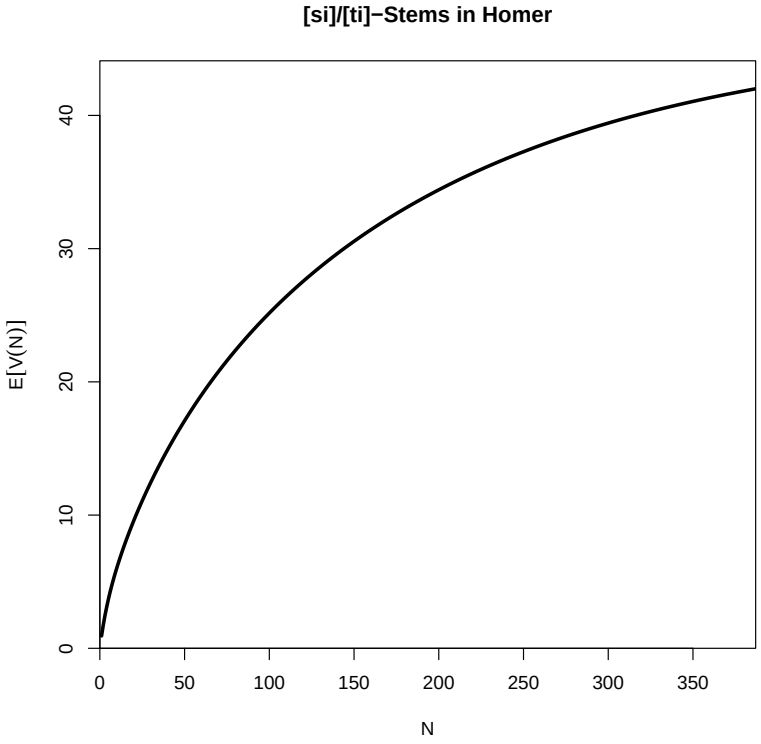



Figure 2.3: Binomially Interpolated Vocabulary Growth Curve for [si-]/[ti]-Stems in Homer

The mathematical basis for $\mathscr{P}$ makes clearer the importance of hapax legomena: the faster the rate at which new types belonging to a morphological category enter the item-sample (i.e., the corpus), the more productive that category must be. Since that rate of growth in types belonging to the category is the slope of the growth curve, which $n_1/N$ expresses, then $\mathscr{P}$ is a valid measure of the growth rate. See Baayen 1992: 112–9, and especially Baayen 1989: 99–107 for a fuller mathematical explication of $\mathscr{P}$. In particular, Baayen proves that the ratio $n_1/N$ indeed approximates the slope of the tangent to a VGC at a given $N$. Granted then that the ratio $n_1/N$ approximates rate of vocabulary growth, the measure $\mathscr{P}$ is then "in a very real sense the probability that new types will be encountered when the item sample is increased."

Consider that, for a comparatively unproductive category, such as English nouns in *-th*, the number of new types will quickly be exhausted in an item-sample (corpus).

[10] Vocabulary growth curves in Chapters 5 and 6 will always be given in binomially interpolated, rather than empirical form. The reason is mainly practical, since constructing an empirical vocabulary growth curve requires a means of efficiently accumulating the number of types and tokens belonging to a category as one scans through a corpus from beginning to end, and that requires each token to be appropriately tagged or otherwise identifiable in some fashion.

The slope of the growth curve for such a formation then will rapidly approach $0$, i.e., become asymptotic. Conversely, for a productive category, such as English nouns in *-ness*, the slope of the growth curve will remain constantly positive (i.e., the rate of increase will stabilize at a value above zero), since its capacity to add new members is practically unbounded. That is to say, it may be true that $\lim_{N\to\infty} V(\textit{-ness}) = \infty$, while the same limit of *-th* will be some finite integer. This distinction between productive and unproductive categories also plays into the measure $\mathscr{I}$ treated below.

These considerations are empirically borne out in an examination of the English derivational suffixes *-th*, *-ity*, and *-ness* based on the text of Herman Melville's *Moby-Dick* (published in 1851).[11] In the unlemmatized text that I employ here, there are $215994$ tokens ($N = 215994$) and $20531$ distinct surface types ($V = 20531$). Even at 150 years of remove from Melville, I (and as repeatedly reported in the literature, speakers of English generally) share the same intuition concerning the productivity of these three derivational suffixes as the quantitative data from *Moby-Dick* indicate: *-ness* is decidedly productive, *-ity* appreciably less so than *-ness*, and nominal-deriving *-th* is wholly unproductive. Table 2.1 gives the token frequency ($N$), type frequency ($V$), number of hapax legomena ($n_1$), and the derived measure $\mathscr{P}$ (= $n_1/N$) for each of these three suffixes.[12] As expected $\mathscr{P}$ for *-ness* is more than twice as great as $\mathscr{P}$ for *-ity*, while the absence of any hapax legomena whatsoever among forms with *-th* marks the suffix as entirely unproductive, in accordance with intuition.

| Category | $V$ | $N$ | $n_1$ | $\mathscr{P}$ |
|---|---|---|---|---|
| *-ness* | 270 | 565 | 186 | 0.3292035 |
| *-ity* | 153 | 499 | 70 | 0.1402806 |
| *-th* | 15 | 205 | 0 | 0 |

Table 2.1: Frequency Statistics and $\mathscr{P}$ for English *-ness*, *-ity*, and *-th* in *Moby-Dick*

The sharp differences in the expected behavior of the growth curves expected for these categories clearly emerges when the binomially interpolated VGCs based on the above data[13] are plotted, as in Figure 2.4.[14] As described above, the VGC for *-th* quickly begins to flatten, approaching a slope of $0$, beyond which point no new types are expected.[15] Conversely, *-ness* is growing in types rapidly, and, at least in this limited corpus, gives

[11] The electronic text of *Moby-Dick* used here is included in the `languageR` (Baayen 2013 [2007]) `R` package.

[12] The fact that *Moby-Dick*, as a corpus, is relatively small (in number of tokens, it is not substantially larger than the Homeric epics or the *Ṛgveda*, which will be used in the study of Greek and Sanskrit morphology here), but that the quantitative differences in productivity between these three suffixes are clear, is reassuring; it indicates that corpora of even just $\sim 200000$ tokens may be adequate for capturing differences in productivity. It is worth noting, however, that *-ness* and *-ity* appear very productive overall, with much higher $\mathscr{P}$ than most categories that will be examined in Ancient Greek or Vedic Sanskrit.

[13] Strictly speaking, creating those VGCs requires the full frequency spectrum (see 2.3 below) for a category.

[14] Note that the curves are of different length simply because there are fewer tokens of *-ity* and *-th* than of *-ness*.

[15] In the data underlying this curve, the number of hapax legomena falls below 1 where $N = 162$.

not the slightest indication of soon approaching its "true" number of types. *-ity*, for its part, is plainly not as robust as *-ness*, but very different from the enfeebled *-th*. Figure 2.4 thus gives a clear visual representation of the quantitative differences in productivity between these three suffixes expressed by their values for $\mathscr{P}$ given in Table 2.1.

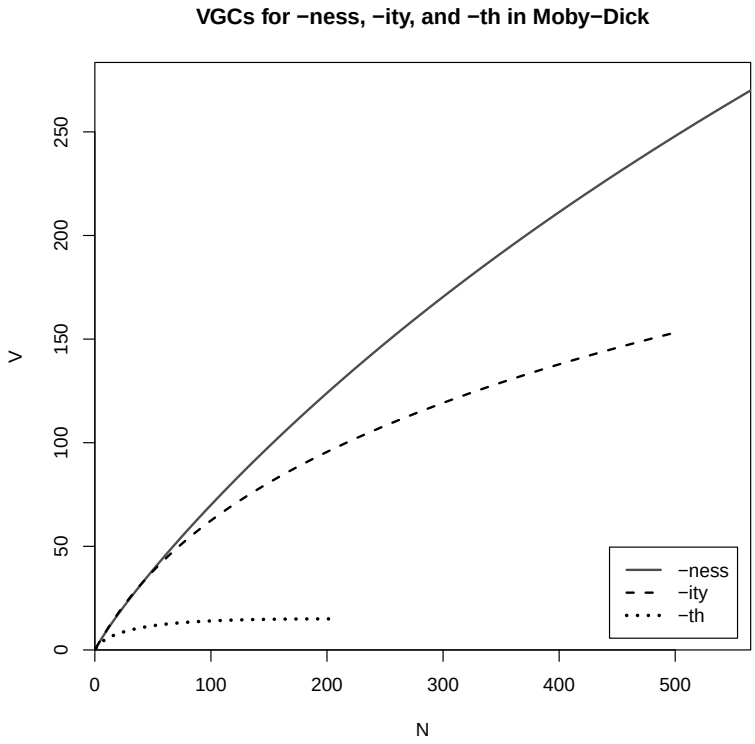



Figure 2.4: Vocabulary Growth Curves for the English Suffixes *-ness*, *-ity*, and *-th* in *Moby-Dick*

2. $\mathscr{U}$ : pragmatic usefulness, "the extent to which a rule can appropriately be put to use within the socio-cultural matrix" is equivalent to the number of types ($V$) that instantiate that rule. Hence, $\mathscr{U} = V$. With respect to a given language sample, the pragmatic usefulness of rules may also vary, depending upon the topical nature of the sample.[16] For instance, Baayen (1989: 24–5) attributes the low number of types in the Eindhoven corpus with the pejorative suffix *-erd* (e.g., *bangerd* 'coward' from *bang* 'afraid') in samples consisting of written Dutch to the sense that *-erd* is inappropriate for written registers. In effect, pragmatic factors may play some role in reducing the strict productivity of a formation. If pragmatic forces drive the usage of

[16] In the same way, the token frequency of specific words is an indicator of the pragmatic usefulness of that word in the context of a particular "socio-cultural matrix" or language sample. Hence, that forms of names of significant deities in the Vedic pantheon (especially in the vocative), have high token frequency is the RV is a consequence of the text's pragmatic function.

a formation to a very low degree, then the process of generating that formation may become unlearnable, and the formation may die away. On the other hand, a low degree of pragmatic usefulness may result in a relatively high ratio of hapax legomena to tokens, and thereby indicate to speakers that a word-formation process belies much untapped potential.

3. $\mathscr{I}$: pragmatic potentiality, "the potentiality of a word-formation rule as it manifests itself within the socio-cultural matrix" is measured as the ratio of the total number of types in the population (*S*, i.e., every potential word belonging to a particular formation formed by every speaker of a language) to the number of types belonging to a particular formation; the measure would indicate how much growth potential a category has. Formally:

$$\mathscr{I} = \frac{S}{V} \tag{2.2}$$

This measure simply inverts the ratio proposed by Aronoff of sample types *V* to population types *S*, which would indicate what proportion of possible items in fact exist. In principle, unproductive categories should have very low values of $\mathscr{I}$, because the set of potential words will be coextensive with the set of actual words: $\mathscr{I}$ will approach a value of 1. Meanwhile, productive categories should have high values of $\mathscr{I}$, in principle approaching infinity. $\mathscr{I}$ thus theoretically complements $\mathscr{P}$, by providing a measure based on competence (the population types) alongside a measure based on performance (the sample types).

In subsequent work, Baayen has largely left the measures $\mathscr{U}$ and $\mathscr{I}$ aside. Although Baayen (1989) recognized techniques for estimating $\hat{S}$, he found the results to be unreliable. Moreover, the cases in which $\mathscr{P}$ and $\mathscr{I}$ will produce different relative estimates for productivity will be just those cases in which a category contains a substantial number of low-frequency types (thus indicative of more unseen or unrealized types) but a very large number of tokens, leading to a small value for $\mathscr{P}$. $\mathscr{U}$, on the other hand, evidently could still contribute towards explaining certain morphological distributions, namely, where pragmatic factors are identifiable; Hay and Baayen (2002) and Hay and Baayen (2003) do continue to take account of type frequency ($V = \mathscr{U}$). Baayen (1993), meanwhile, introduced a further metric:

- $\mathscr{P}*$: global productivity, also called the "hapax conditioned degree of productivity" (Hay and Baayen 2002: 18), is measured as the ratio of the hapax legomena belonging to a morphological category ($n_1$) to the total number of hapaxes in a corpus ($h_1$).[17] $\mathscr{P}*$ thus measures the contribution that a particular morphological category makes

[17] Note that, since the denominator of this ratio, the total number of hapaxes in the corpus ($h_1$) remains constant for all data from a single corpus, one can compare $\mathscr{P}*$ for different processes within a corpus using only the number of hapaxes belonging to those processes ($n_1$), rather than fully calculating the ratio. Baayen normally follows this procedure, simply treating $n_1$ for a process as $\mathscr{P}*$.

to the total growth rate of the vocabulary in a corpus. If the rate of growth for $V$ in the entire corpus is measured by $\frac{h_1}{N}$ for the entire corpus, then the proportion of $n_1$ in the a given category then indicates the amount of vocabulary growth that that category explains. $\mathscr{P}*$ evidently does not strictly measure productivity itself, since for instance, there are 294 hapax legomena simplex nouns in the Eindhoven corpus (more than for the suffix *-heid*), although simplex nouns have a very low $\mathscr{P}$ value. Baayen (1993: 194) therefore suggests that "$\mathscr{P}$ and $\mathscr{P}*$ are complementary measures, the primary use of $\mathscr{P}$ being to distinguish between productive and unproductive processes as such, $\mathscr{P}*$ being especially suited to ranking productive affixes." This role for $\mathscr{P}*$ emerges because hapax legomena are more likely to be subject to the application of a word-formation process (in both production and perception), rather than whole-word lexical access (see Chapter 3), and thereby psychologically activate that process. The more cognitively active that a word-formation process is, the more probable it is that a speaker will employ it.

The measures $\mathscr{P}$ and $\mathscr{P}*$ have obtained some degree of purchase in the analysis of morphological productivity, as their usage in the literature indicates: Bauer (2001: Ch. 6), Bolozky (1999), Plag (1999), and Säily (2008) all make use of these two measures; the studies of Hay and Baayen (2002) and Hay and Baayen (2003) are concerned with the further psycholinguistic and statistical validation of those two measures (discussed specifically under 3.4).

Two issues remain at this point: one is the estimation of the population for a category $S$, which is necessary to measure $\mathscr{I}$ at all; the other is how these measures, once obtained are to be interpreted. I take up these issues before introducing analogical modeling techniques as another potential window into productivity. An understanding of how to estimate $\hat{S}$ is predicated upon some familiarity with the properties of word frequency distributions. Indeed, because research on the description of word frequency distributions generally theoretically undergirds the productivity measures here, I use the following section as a brief introduction, leading up to the method of calculating $\mathscr{I}$.

### 2.2.3 Word Frequency Distributions

From a statistical point of view, the distributions of words in natural language texts, in corpora ranging in size from mere thousands up to billions of words, exhibit a peculiar, though well-known property: very many items cluster in low frequency ranges (i.e., there are many distinct types with a token frequency of $1$ or $2$), but rather few in higher frequency ranges (e.g., perhaps only one type with a token frequency of $3426$).[18] If words from a sample are ranked from most frequent (frequency rank $1$) to least frequent, the difference in token frequencies at the highest frequency ranks shows an exponential change, a linear change in middle frequency ranks, and long rightward tails at the lowest frequency ranks. Zipf (1935)

[18] For deeper, more detailed, and mathematically more complex discussion of all the notions treated in this section, see generally Baayen 1989: Chapters 5–7 and Baayen 2001.

was among the earliest to explicitly describe this distribution in language.[19] Consider Table 2.2, which gives the twenty most frequent (surface) word forms in Melville's *Moby-Dick* (token frequency $N = 215994$, type frequency $V = 20531$), where $z$ is the Zipf Rank and $f_z(z, N)$ is the resulting token frequency from a function that accepts a Zipf rank $z$ and some token sample size $N$; in this case, $N = 215994$, the total number of tokens in my text of *Moby-Dick*.

Table 2.2: Frequency Rank of the twenty highest-ranked surface forms in Moby Dick, ordered in decreasing frequency.

| $z$ | $f_z(z, N)$ | word | $f_z(z, N)$ | $N$ | word |
|---|---|---|---|---|---|
| 1 | 13717 | *the* | 11 | 1732 | *'s* |
| 2 | 6512 | *of* | 12 | 1695 | *is* |
| 3 | 6008 | *and* | 13 | 1661 | *he* |
| 4 | 4551 | *a* | 14 | 1659 | *with* |
| 5 | 4514 | *to* | 15 | 1632 | *was* |
| 6 | 3908 | *in* | 16 | 1620 | *as* |
| 7 | 2982 | *that* | 17 | 1446 | *all* |
| 8 | 2457 | *his* | 18 | 1414 | *for* |
| 9 | 2209 | *it* | 19 | 1280 | *this* |
| 10 | 2122 | *I* | 20 | 1230 | *at* |

Meanwhile, all words in this text with frequency ranks 10250 to 20531 have a token frequency of 1. Graphically, plotting the log-transformed[20] token frequencies $N$ against log-transformed frequency ranks $z$ shows a reasonably linear fit within much of the middle of the distribution, but an exponential fit at the left edge and the long tails at the right edge (Figure 2.5)

A similarly formed distribution, for instance, applies to the (surface) word types in the saṃhitā text of the *Ṛgveda* (Figure 2.6).[21] Here, I have also included a regression line[22] that

[19] The Zipf distribution is also known as the zeta ($\zeta$) distribution, since it is equivalent to the Riemann zeta function (cf. Baayen 2001: 15); some may know it under the name of the Pareto distribution, due to the Italian economist Vilfredo Pareto who employed it to describe family incomes. One normally speaks of Zipfian distributions where the elements are discrete (like words), but a Pareto distribution where the elements are continuous (Ross 2010: 163–4).

[20] Recall that a logarithm with a given base $b$ of some number $x$ is the exponent $y$ of the base $b$ that produces the number $x$ (e.g., $log_{10}(1000) = 10, log_5(25) = 2$). All logarithmically transformed data here employ the natural logarithm (i.e., $log_e$, where $e \approx 2.718281828$). These transformations serve to compress heavily skewed data.

[21] The number of unique types is somewhat inflated, because the application of sandhi creates more distinct surface forms; moreover, in this highly inflected language, a non-lemmatized text further multiplies distinct forms. Nevertheless, the sort of distribution would clearly remain the same even with these adjustments.

[22] This regression line plots the predicted log frequency (the *dependent* or *response* variable) given an input log rank (the *predictor* variable). The log frequencies are predicted by solving for the a constant term (or *intercept*) and a coefficient(s) by which to multiply the predictor variable(s) using the method of least-squares. See


Figure 2.5: Zipfian distribution in Melville's *Moby-Dick*

relates log frequency as a function of log rank. This makes very clear the good fit of a linear model in the middle frequency range, but clearly does not account for the divergence at high frequency, nor for the large number of low-frequency items.

The issues that affect the frequency distributions of a vocabulary as a whole are of concern to us precisely because they usually apply to subsets of that vocabulary, such as all the lexical items belonging to a given morphological category. Figure 2.3 shows that sigmatic aorists (cf. the examples (4) under Section 1.1 above) in Homer ($N = 6749$, $V = 615$; cf. Section 5.3) likewise assume a Zipfian distribution.[23] Indeed, the Zipfian properties of word frequency may themselves be indicative of (non-)productivity, especially where subsets do not look especially Zipfian. Consider, in comparison, the plot of root aorist frequency ranks ($N = 2710$, $V = 41$; cf. again Section 5.3) in Figure 2.8, again based on data from Homer. The distribution in this case shows a markedly less Zipfian character, in that its rightward skew appears substantially weaker; indeed, the tail at log frequency $1$ appears longer than the tail at log frequency $0$;[24] there are five types at log frequency $1$, but only $2$ at log frequency

Lantz 2013: 169 for an efficient algebraic means of calculating regression coefficients, and Baayen 2008: 82–11; 169–94 or Johnson 2008: 57–67 for good introductions to regression analysis as applied to linguistic data.

[23] On the terms "sigmatic aorist" and "root aorist", see Section 5.1.

[24] Log frequency $0$ is equal to a token frequency of $1$.

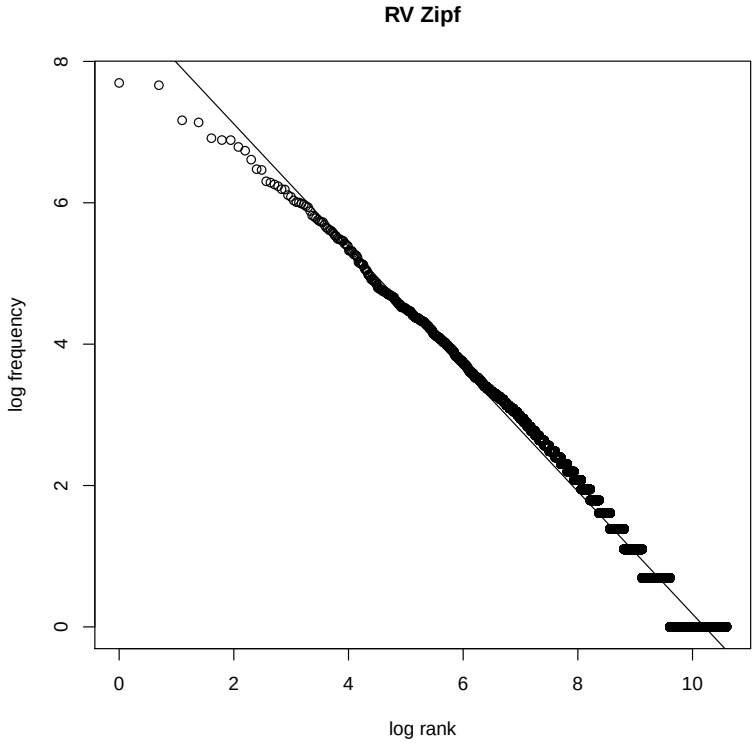



Figure 2.6: Zipfian distribution in the *Ṛgveda*

0. This distributional fact does not fit the usual word frequency distributions of texts, which have a true population much greater than the number of types actually found in the sample, nor of productive categories like English nouns in *-ness*, which have many more types with a log frequency of 0 than at any other log frequency; in parallel, English nouns in *-th* show no items with a log frequency of 0.

The skew in these word frequency distributions can also be represented by way of a **frequency spectrum**, which shows the number of distinct types ($V$) that have a given token frequency ($m$) within a total sample of $N$ tokens (following Baayen 2001: 8). For instance, in the total sample of 215994 tokens ($N$) that constitute the text of *Moby-Dick*, there are 10282 types ($V$) that have a token frequency of 1 ($m$). Formally, the number of types at a given frequency rank is a function of the number of sample tokens $N$ and the frequency rank $m$; for *Moby-Dick* (as a subsample of English in the year $\sim$ 1851):

$$V(1, 215994) = 10282; V(2, 215994) = 3299; etc. \tag{2.3}$$

The complete frequency spectrum for any category (lexical items in a corpus, members of a morphological category, undergoers of a phonological process, etc.) simply relates the token frequency index $m$ to the type frequency $V$ for all token frequency indices. As examples, I give the full frequency spectrums for root aorists occurring in Homer in Table 2.3 and

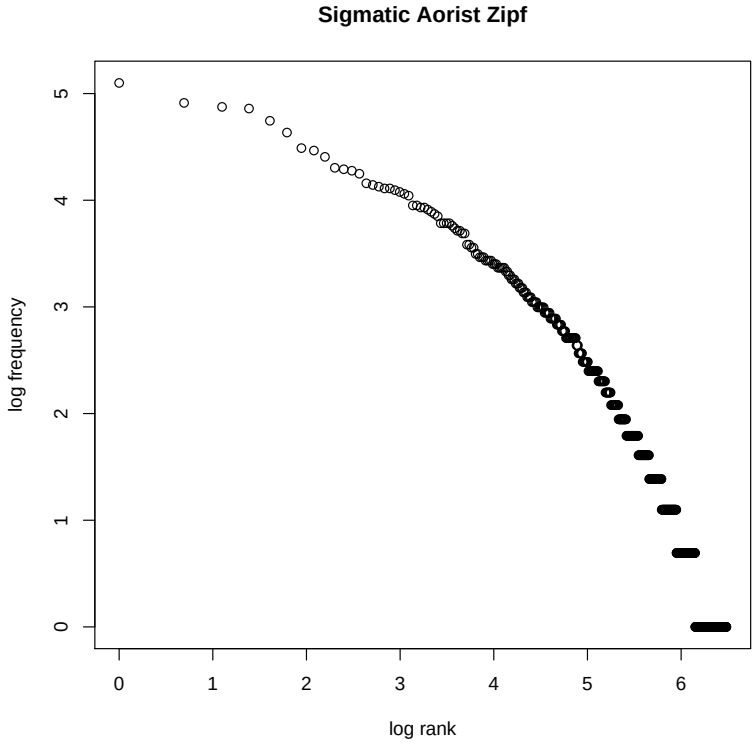



Figure 2.7: Zipfian distribution of Sigmatic Aorists in Homer

for nouns derived with the suffix *-ness* in *Moby-Dick* in Table 2.4. We see that, among the Homeric root aorists, there are two types that occur just once (two *hapax legomena*), three types that occur twice (two *dis legomena*), ..., and one type that occurs $431$ times (with a total of $27$ distinct frequency bands). *-ness* in *Moby-Dick* is instead heavily balanced towards infrequent items: there are $186$ items that occur just once and $45$ that occur just twice, while only single items occur $8$ or more times.

Such frequency spectra represent precisely the inverse of the Zipf frequency rank $z$ for some sample (Baayen 2001: 13–4): for the text of *Moby-Dick*, $f_z(1, N) = 13717$, and in turn $V(13717, N) = 1$ (the lexical item *the*; cf. Table 2.2 above). This property in a different way brings out the same point, and the same problem: typically, most types are represented by very few tokens, but a small number of types represent an inordinate number of tokens. Consider again the line fit to the plot in Figure 2.6. Generally speaking, the heart of the problem is this: the considerable quantity of items at the low end and high ends of the frequency spectrum requires some special treatment. Zipf (1935) (cf. discussion in Baayen 2001: 13–24)

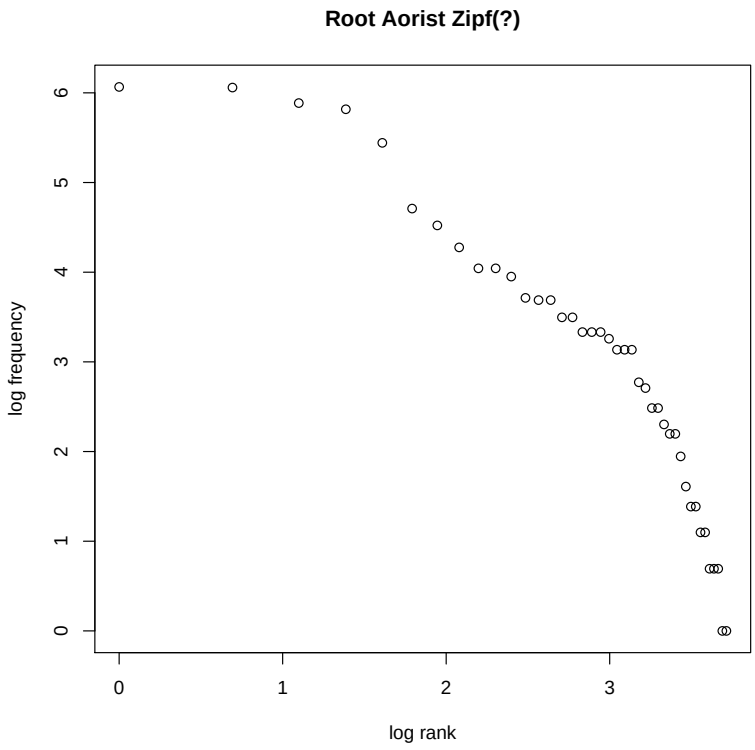



Figure 2.8: Non-Zipfian (?) distribution of Root Aorists in Homer

| $m$ | $V(m,N)$ | $m$ | $V(m,N)$ |
|---|---|---|---|
| 1.00 | 2.00 | 33.00 | 2.00 |
| 2.00 | 3.00 | 40.00 | 2.00 |
| 3.00 | 2.00 | 41.00 | 1.00 |
| 4.00 | 2.00 | 52.00 | 1.00 |
| 5.00 | 1.00 | 57.00 | 2.00 |
| 7.00 | 1.00 | 72.00 | 1.00 |
| 9.00 | 2.00 | 92.00 | 1.00 |
| 10.00 | 1.00 | 111.00 | 1.00 |
| 12.00 | 2.00 | 231.00 | 1.00 |
| 15.00 | 1.00 | 336.00 | 1.00 |
| 16.00 | 1.00 | 360.00 | 1.00 |
| 23.00 | 3.00 | 428.00 | 1.00 |
| 26.00 | 1.00 | 431.00 | 1.00 |
| 28.00 | 3.00 | – | – |

Table 2.3: The Frequency Spectrum of Root Aorists in Homer

| $m$ | $V(m, N)$ |
|---|---|
| 1.00 | 186.00 |
| 2.00 | 45.00 |
| 3.00 | 21.00 |
| 4.00 | 4.00 |
| 5.00 | 3.00 |
| 6.00 | 3.00 |
| 7.00 | 2.00 |
| 8.00 | 1.00 |
| 11.00 | 1.00 |
| 17.00 | 1.00 |
| 28.00 | 1.00 |
| 32.00 | 1.00 |
| 67.00 | 1.00 |

Table 2.4: The Frequency Spectrum of Nouns in *-ness* in *Moby-Dick*.

formulated a power law[25] as an attempt at capturing this distribution:

$$f_z(z, N) = \frac{C}{z^a} \tag{2.4}$$

The number of tokens belonging to an item having a frequency rank $z$ in a total sample size $N$ (i.e., $f_z(z, N)$) is given by a normalizing constant $C$ (which ensures that all of the token frequencies resulting from $f_z(z, N)$ indeed sum up to $N$) divided by that frequency rank $z$ raised to the power of the free parameter $a$. The log-transformed version of this equation, thus reflecting the inverse relation between log frequency and log rank represented clearly in Figures 2.5–7, is **Zipf's Law**, properly speaking:

$$\log f_z(z, N) = \log C - a \log z \tag{2.5}$$

We can also think of Zipf's Law as describing the probability of sampling a given item, which is computed simply as the number of tokens belonging to that item, divided by the total sample size $N$, thus $\frac{f_z(z,N)}{N}$. If we say that

$$\frac{f_z(z, N)}{N} = \frac{C}{z^a} \tag{2.6}$$

the function of the normalizing constant $C$ then becomes to ensure that the sums of the individual item probabilities indeed sum to $1$.

[25] A power law is a function whereby one quantity can be described by the exponentiation of another quantity. In this case, then, the number of items at some frequency rank $z$ is determined in part through the exponentiation of that frequency rank value ($z^a$).

Furthermore, since, as shown above, the outputs of $f_z(z, N)$ and $V(m, N)$ are their respective inverses, Zipf's Law can also be stated in terms of $V(m, N)$:

$$V(m, N) = \frac{C}{m(m+1)} \tag{2.7}$$

Under the assumption that, in an ideally Zipfian distribution, events occurring just once make up half of the types, the normalizing constant $C$ can be set equal to the number of types, $V(N)$. This restatement thus emphasizes enormous overabundance of types having the lowest possible discrete frequency (i.e., $1$; cf. Evert 2004: 14). Nevertheless, it remains the case that the frequencies and probabilities of the lowest-ranked items are still captured poorly by Zipf's Law: the power law extends to fit the few very frequent items, but is unable to handle the long rightward tails, i.e., the many very infrequent items that occur in the sample. One (of many) proposed adjustments to Zipf's Law intended to address this problem is simply to add another free parameter, call it $b$, to the model, as Mandelbrot (1953) proposed:

$$f_z(z, N) = \frac{C}{(z+b)^a} \tag{2.8}$$

This further parameter $b$ has the effect of maintaining the output of the function $f_z(z, N)$ at approximately $1$ for large $z$, which is desirable for elements that can only assume discrete values (like words). Equation 2.6 is then appropriately called the Zipf-Mandelbrot Law; it is central to the discussion on the calculation of "pragmatic potentiality" ($\mathscr{I}$) in the following section, 2.2.3.1.

Even observationally, the figures 2.5–7 appear to conform to a LARGE-NUMBER OF RARE EVENTS (LNRE) distribution – very many (as much as half) of the distinct events take the minimum discrete probability (i.e., $1/N$). Baayen (1989: 96) defines such a LNRE distribution as having two properties (following Khalmadze 1987: 12):

**Definition 1**

*A sequence of types* $\{\mathscr{V}^N\}$ *is called an LNRE sequence if*

$$\lim_{N\to\infty} \frac{E[n_1]}{E[V]} > 0 \; and \; \lim_{N\to\infty} E[V] = \infty$$

In other words: as the number of tokens being sampled increases towards infinity ($\lim_{N\to\infty}$), two conditions should hold: the ratio of the expectation of the number of events that occur only one time ($E[n_1]$) to the expectation of the number of types ($E[V]$) should be greater than $0$, while the expectation of the number of types itself should be infinite. Strictly speaking, these conditions probably do not hold for human language, and pragmatically speaking, at the level of individual speakers, it is doubtful that any speaker can generate a truly infinite number of lexical items (or would choose to employ all of them). Baayen (1989: 98) argues that LNRE models remain appropriate for describing word frequency distributions because the conditions under which a chi-squared ($\chi^2$) goodness-of-fit test can appropriately applied to evaluate the differences in word frequency samples are not usually met; specifically,

the chi-squared distribution does not become symmetrical at large samples of types (this is normally expected where $V > 100$). Since this property holds of true LNRE sequences, to treat word frequency distributions as LNRE sequences will be adequate.

From the preceding treatment, I can now summarize three major points concerning word frequency distributions that should be borne in mind going forward.

- The relationship between log rank and log token frequency in some sample of natural language, both of all items or of items exhibiting a particular property (such as a specific derivational suffix) is approximately described by Zipf's Law and the Zipf-Mandelbrot Law.
- An inverse relationship exists between frequency rank $z$ and the number of types occurring with some token frequency $V(m, N)$ as described by a frequency spectrum.
- Samples drawn from natural language that exhibit a frequency distribution markedly different from distributions that can be approximated by Zipf's Law (specifically, having few low frequency events) seem to parallel unproductivity.
- Word frequencies can be said to belong to the class of large-number of rare event distributions; informally put, many different individual events that each themselves take up a very small proportion of the probability mass together constitute a considerable proportion thereof.
- Natural language samples that deviate markedly from a Zipfian distribution, in particular, the absence of a long right tail (i.e., having few very low frequency events) may reflect populations from which most types have already been sampled (cf. Figure 2.8 above).

#### 2.2.3.1 Calculation of $\mathscr{I}$

Besides $\mathscr{P}$, which is easily calculable from frequencies from corpus data ($n_1$ and $N$ of a category), to be able to calculate the potential productivity ($\mathscr{I}$) would be ideal. The fundamental difficulty that besets an attempt at calculating $\mathscr{I}$ is that it cannot be solved for from the direct empirical examination of corpus data alone (unlike $\mathscr{P}$, $\mathscr{P}^*$, and $\mathscr{U}$). The problem lies in obtaining a confident estimate of $S$, the total population size (the overall number of types), belonging to a category. To be explicit: the complete set of *all* attested types and *potential unseen* neologisms is $S$:

$$S = \lim_{N \to \infty} V \tag{2.9}$$

In principle, the solution to the estimation of $S$ is straightforward: if we have a mathematical model that can fit the empirical growth curve (like those in Figures 2.1 and 2.2), then one need only to extrapolate beyond the number of observed tokens to see how the number of types should correspondingly increase. The point at which the growth curve becomes asymptotic, i.e., fails to increase any further, will be the estimated size of the population of

types. Zipf's Law, and the Zipf-Mandelbrot Law discussed above, as possible models of word frequency, are able to obtain the necessary extrapolations.

While Baayen (1989: 145–87) is aware of various methods for estimating $S$, he is skeptical how well the Zipf-Mandelbrot Law and other revisions of Zipf's Law (e.g., Waring-Herdan-Müller) fit the empirical distributions for a number of morphological categories. Namely, the Zipf-Mandelbrot density function will lead to an infinite population, which is unrealistic for natural languages; Evert (2004), on the other hand, finds that *finite* implementations of the Zipf-Mandelbrot Law to be mathematically elegant and empirically accurate for large corpus datasets. Evert therefore formulates a finite Zipf-Mandelbrot model (fZM) by introducing a lower cutoff point for type density. Evert (2004: 5–7) provides a density function for the Zipf-Mandelbrot law that can be used to derive an LNRE model.

More importantly, the comparisons of vocabulary size extrapolations tested for both Zipf-Mandelbrot and finite Zipf-Mandelbrot models in Baroni and Evert 2005 for German derivational suffixes with relatively small category $N$ shows that fZM extrapolations, in Evert's implementation, are very reliable. Hence, the estimates for $S$ obtained by extrapolation using fZM models can be largely trusted. Zeldes (2012: 76–85) also discusses Evert's work and accepts its conclusions. In addition, $\chi^2$ goodness-of-fit tests can be used to evaluate the validity of any single model. I therefore use the implementations of the fZM model available in Evert and Baroni 2008 to estimate $S$. As an example, Figure 2.9 shows the extrapolated vocabulary growth curve of *Moby-Dick* out to $100000$ tokens (the black line up to $N = 215994$ is the empirical growth curve); at this size, the growth curve is not yet asymptotic. The fZM model that underlies the extrapolated growth curve estimates $S$ to be $43389$; the attested $V$ in the text of *Moby-Dick* is $20531$, and hence $\mathscr{I}$ is $2.113341$. This number indicates that, as one would expect, a substantially larger population of word types, more than twice the number actually occurring in *Moby-Dick*, can be estimated to exist and have been available in the population from which the forms occurring in *Moby-Dick* were drawn.

#### 2.2.3.2 Comparing Productivity Measures

Now that I have described several measures of productivity in their strictly mathematical and statistical terms, it is appropriate to ask a two-fold question: 1) what do these measures mean theoretically, i.e., what implications do they hold for the study of morphology in general and historical morphology specifically; and 2) how are the results from such measures to be appropriately deployed and compared?

To my mind, the most significant contribution that these measures make is to offer insight into the status of mophologically complex words: words belonging to unproductive categories are rather more like morphological simplexes than words belonging to productive categories. Proof of this point is inevitably bound up with psycholinguistic evidence, which will be the topic of Chapter 3. From the point of view of the historical linguist, the ability to concretely measure productivity can alert us to a level of language change that might ordinarily be too subtle to be discerned; in being able to observe the phenomenon (change in degree of productivity) directly, we are then in a position to attempt explana-

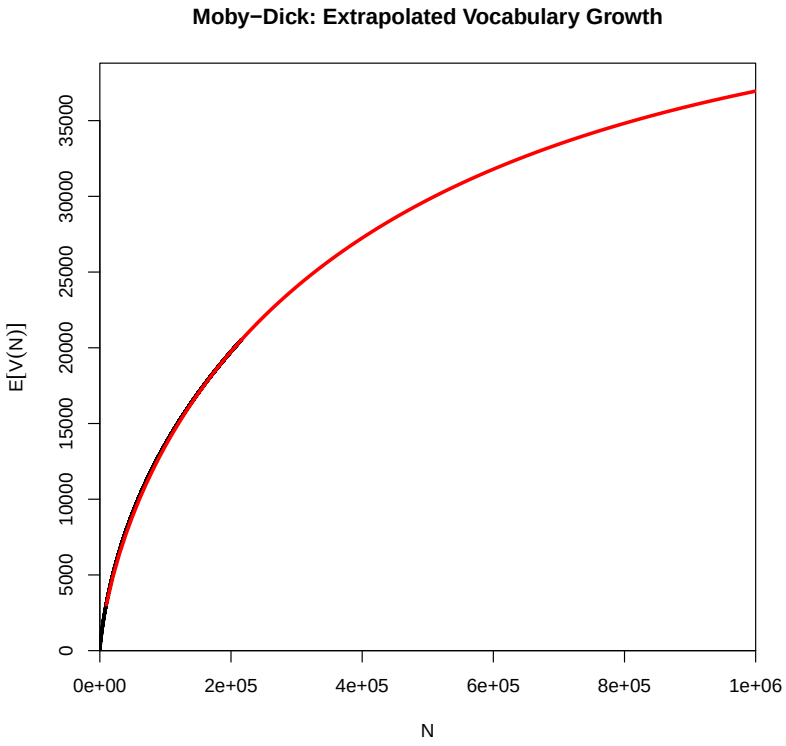



Figure 2.9: Extrapolated Vocabulary Growth Curve for *Moby-Dick*, up to $100000$ Tokens

tions. Furthermore, where the productivity of a category may impact other phenomena, we gain a potential explanatory mechanism.

For the moment, however, I wish to focus on the second question. What information beyond bald statistics is being communicated when we say that a category's $\mathscr{P}$ is measured as $0.000735$ or $0.0284$? It may seem self-evident to say that the latter category (granting that the data derive from the same corpus) is more productive than the former by a substantial margin. However, several further examinations of the measure $\mathscr{P}$ in particular (especially Gaeta and Ricca 2006, Gaeta 2007, and Zeldes 2012: Ch. 2) have revealed some potential issues with the direct comparison of values for $\mathscr{P}$ based on the category-specific values of $N$ and $n_1$ obtained from a given corpus.

First, note that the functions $\mathscr{P}(N)$ and $\mathscr{P}*(N)$ are both decreasing non-monotonic functions – the more tokens that are sampled, the smaller the values calculated for $\mathscr{P}$ will be, because fewer and fewer hapax legomena proportional to tokens will appear (cf. Baayen and Lieber 1991: 837). Gaeta and Ricca (2006) empirically demonstrate that, in fact, $\mathscr{P}$ approaches $0$ as $N$ approaches $\infty$, even for affixes of very different productivities.[26] $\mathscr{I}$ will likewise approach $1$ as $N$ approaches $\infty$ for such affixes. As a consequence, even in cases of

[26]Namely, Gaeta and Ricca show that $\mathscr{P}$ grows closer to $0$ with larger and larger $N$, but for none of the Italian affixes that they discuss does $\mathscr{P}$ actually reach $0$.

intuitively productive derivational processes (English *-ness*, Dutch *-heid*, Italian *-zione*), as ever larger numbers of tokens are sampled, the value of $\mathscr{P}$ will decrease.[27]

The very sensitivity of computed values for $\mathscr{P}$ to the size of the denominator, $N$, leads Gaeta and Ricca (2006: 62–3) to argue that $\mathscr{P}$ is distorted as a measure of productivity, even for data gathered from the same corpus. Their proposal is instead to compare $\mathscr{P}(N)$ at equal sizes of $N$ for each morphological category being compared; practically speaking, one can either gather tokens through a corpus up to a certain limit, or estimate the growth curve from $0$ to the limited $N$ through binomial interpolation (cf. Figure 2.3 above); they call this procedure a *variable corpus* approach. As a further proof of concept, Gaeta (2007) shows that, for some Italian inflectional categories (e.g., 3.sg.pres. vs. 3.pl.pres.), which one might expect to have a theoretically equal degree of productivity, $\mathscr{P}(N)$ is substantially different at different $N$, but nearly identical when $N$ is the same.[28]

On the one hand, while it may be true that the $\mathscr{P}$ score for a category is overestimated for categories having a low $N$ in a corpus, logically, a corpus is intended as a sample of language as a whole, so higher token frequency categories in a corpus should likewise reflect higher token frequencies in the language as a whole. Indeed, in any corpus that has a total $N$ substantially less than the $N$ to which a speaker has been exposed, the $\mathscr{P}$ for virtually all categories would be an overestimate. Furthermore, we have no reason to believe that the rate at which new tokens are added to a speaker's exemplars is any different for a corpus; while tokens may be occur at somewhat irregular intervals (i.e., be underdispersed), the growth of $N$ is roughly constant. Thus, it seems unrealistic to me to compare $\mathscr{P}$ between two categories by setting $N$ at the $N$ of the category with fewer tokens; perhaps more realistic would be to proportionally reduce the sample $N$ for the categories under examination.

To take a simple example, we know that, in the entire Homeric corpus ($\sim N = 200000$), [si-]/[ti]-stems (e.g., [brɔ̂ːsis] 'food') have $387$ tokens, while [tu]-stems (e.g., [edɛːtús] 'food') have only $196$ tokens. Assuming (unrealistically) that tokens are evenly distributed, a corpus of half the size would contain half as many tokens, thus $194$ and $88$, respectively; the number of hapax legomena, however, will not decrease linearly, but must be determined through binomial interpolation to that smaller sample size. With at $N = 387$ and $N = 196$, $\mathscr{P} =$

[27] This result is both a mathematical consequence of how $\mathscr{P}$ is defined, and as a result of the fact that more useful lexical items that belong to a category will continue to accumulate tokens, while the introduction of new types becomes ever more infrequent. Nonetheless, it is conceivable that, at some very large token sample that exhausts all lexemes that are simply rare rather than genuine neologisms, the true neologisms might begin to appear in sample at a regular rate (say, once every $10000$ tokens of the category). If so, then then $\mathscr{P}$ for our hypothetical category would not be $0$ as $N$ approaches $\infty$, but would be that regular rate of neologism appearance, $0.0001$. Hence, I believe that $\mathscr{P}$, in the long run, would either appear as $0$, for truly unproductive categories (no new hapax legomena ever appear), or the *true* degree of productivity for a process, for truly productive categories. Yet, in the latter circumstance, the long-run value for $\mathscr{I}$ would be $> 1$, which entails that $V = \infty$ for the category, and that $\mathscr{I}$ itself $= \infty$; for productive categories, one would then wish to extrapolate vocabulate growth with a non-finite Zipf-Mandelbrot model, but one would necessarily have to know in advance whether $\mathscr{P}$, in the long run, is greater than $0$ or reaches $0$, leaving a conundrum.

[28] The other possibility is that these corpus-based procedures for evaluating productivity are not very meaningful for obligatory, contextually determined inflection, such as person and number marking on Italian verbs. Rather, the productivity of such inflectional categories would be simply absolute, deployed whenever necessitated by the grammar.

.023 and $\mathscr{P} = .051$. At the reduced token sizes of $N = 194$ and $N = 88$, the $\mathscr{P}$ for [tu]-stems still exceeds that for [-si]/[-ti]-stems ($0.08436066$ vs. $0.06845366$); only with an overall token sample of somewhat greater than $1/4$ that of the Homeric corpus or less does the $\mathscr{P}$ for [si-]/[ti]-stems exceed that of [tu]-stems. Such reductions would seem patently to amount to an unacceptable loss of data; at best, the procedure of $N$-reduction may replace one estimation problem with another.

One might hypothesize that speakers attend to not only the the token frequencies of categories and the number of novel forms belonging to that category that they produce and process, but to the overall number of tokens of the language that they have encountered, and correspondingly adjust the productivities of the categories. However, in many cases, there may exist some minimum sample of tokens in which the $\mathscr{P}$ scores for two categories are reversed or are much closer than the intuitions of native speakers would allow. For instance, if we sample but $10000$ tokens from the text of *Moby-Dick*, nouns in *-ness* have the following statistics ($N = 13, V = 9, n_1 = 5, \mathscr{P} = .38$), while nouns in *-ity* have the following ($N = 16, V = 10, n_1 = 6, \mathscr{P} = .375$). In this small sample, the two categories are almost indistinguishable from one another. At the sample size of the entire text of *Moby-Dick*, however, the relations of the $\mathscr{P}$ scores have assumed their intuitively expected relationship: $\mathscr{P}(-ness) = .33$; $\mathscr{P}(-ity) = .14$.

Given these considerations, whenever we compare the productivity indices of two categories, $A$ and $B$, if $N(A) > N(B)$ and $\mathscr{P}(A) > \mathscr{P}(B)$, then we can be certain that $A$ is more productive than $B$; reducing the sample of $N(A)$ to equal $N(B)$ would only serve to enlarge the gulf in the estimate of $\hat{\mathscr{P}}$ for categories $A$ and $B$. In the case that $N(A) < N(B)$, but $\mathscr{P}(A) > \mathscr{P}(B)$, the results of measurement with $\mathscr{P}(N)$ may be inadequate. However, the additional availability of $\mathscr{I}$ and $\mathscr{P}*$ should serve to clarify the differences in productivity in such circumstances.

I conclude, therefore, that to reduce category token samples for the purposes of comparison in fact destroys data on the genuine frequency relations that obtain in the language. The fact that Gaeta's variable corpus approach renders similar productivity values for contextually determined inflection may indeed serve as evidence for the fact that such obligatory inflection, which has no competing alternate modes of expression in the language, holds at a productivity of $100\%$ for all the relevant morphemes. Despite Zeldes (2012: 65–83), however, I do not believe that the same procedure ought to be applied in comparing derivational categories: the calculated productivity values for the maximum sample of tokens obtained is what should be employed, at the risk of considering the categories at sample sizes that improperly overestimate the relative productivity of one category with respect to another.

Thus far, I have shown that a few easily calculable and intuitively reasonable corpus-based measures of productivity are available to us; their derivation and properties follow fairly straightforwardly from word frequency distributions more generally. These measures, moreover, provide some quantitative grounding and offer effective synchronic snapshots (however broadly or narrowly "synchrony" may be construed, cf. 1.4.1 above) of productivity. Granting the basic reliability of these measures (see Chapter 3 for some further validation), one can make more informed decisions concerning the antiquity of specific forms. In partic-

ular, to find that a form exists in an unproductive category in any of our compared languages should suggest that the form belongs to an older grammar than the one for which the productivity measurement has been taken.

## 2.3 Modeling Morphological Change: Productivity as Analogy

The other side of recognizing that unproductive categories are largely populated by forms with some degree of archaism is to wonder: what are these categories doing, where are they going, and why? Although the degrees of productivity that measures such as $\mathscr{I}$ and $\mathscr{P}$ furnish make certain predictions inherent in their mathematics, namely, that the productive categories can give rise to many more as yet unseen types, they do not make any explicit claims about the future and the past of that category.

I believe, therefore, that examining corpus word frequency alone cannot, in itself, provide answers to the most desirable questions of all: why do certain forms fall into certain categories, how can we predict what a novel form will be like, and what is the likelihood that a novel form, given certain conditions, will belong to a given category? Baayen's $\mathscr{P}$ measure indicates, in absolute terms based on available data, how likely it is to encounter novel forms belonging to a given category. For instance, based on the results presented in Chapter 5, one can easily say that a Homerid of the 8th c. BCE would have had very strong evidence to permit the creation of new sigmatic aorists, but, conversely, would almost certainly not invent a novel root aorist, and likely even had difficulty recalling and reproducing some root aorists to which he might have been exposed. In the case of truly moribund categories, it may be simple enough to say that the category is closed; yet, if moribund categories are subject to attrition, then we would like to understand which forms on which grounds are most susceptible to transformation, replacement, or loss.

To undertake such tasks requires models that have predictive power. Baayen's measures do not make predictions in any direct way; they merely state whether we are likely to encounter novel instances of a category or not. But it is entirely possible that some relevant factors are being left out of consideration: $\mathscr{P}$ just tells us, in summary fashion, among all the possibilities for a word, whether it will fall into a given bin – its predictive power is very weak, even if the measure itself is very useful. $\mathscr{P}$ is really a sort of summary measure that grossly indicates a baseline likelihood for encountering new types, and in this sense, it essentially satisfies the definition of productivity offered in Chapter 1. Imagine, though, that the existence of a neuter *s*-stem noun in Greek is a good predictor of whether a verb derived from the same root should build a sigmatic aorist. This kind of morphological relationship looks irreducible, whereas a broader measure like $\mathscr{P}$ may break down into, or be derivative of, the quantity and force of the various morphological relations that hold with respect to that category. $\mathscr{P}$ can tell us, with respect to some baseline, how likely it is that we encounter further sigmatic aorists or neuter *s*-stems, but says nothing about the potential relationship that might hold between those categories.

The further questions that interest us thus require **models** that can evaluate the possible relations and make predictions based on linguistic data. Diachrony allows for the opportu-

nity to explicitly test and evaluate these predictions, moreover: to what extent are the seeds later attested changes already sown? Diachronically speaking, changes may occur where the model has weak **recall**, in the technical sense that it fails to reproduce data to which it has been exposed. For example, if Homer attests one type of aorist to a given root, but the New Testament another, why was that particular aorist subject to category change (but not others, which remain stable across the intervening eight hundred years)?

Models are not wizardry (though their outputs may, at times, be difficult to interpret): in order for analysis to function or make reasonable predictions, it must be grounded in 1) accurate data collection and classification (e.g., sigmatic aorist forms should not be classified as thematic aorists)[29], and 2) selection of theoretically motivated parameters (e.g., that token frequency might impact accent assignment, rather than the presence of a racoon in the speaker's visual field). All of these models and techniques presented here are not a substitute for the work of the linguist, but rather help to synthesize large amounts of data, and to bring the generalizations that may hold in that data down to a human scale. These techniques give us a means of evaluating *the extent to which* a given parameter is important (or relevant at all); this also means that datapoints that would stand as outright "exceptions" in an analysis based on just one or two parameters might fall out from the inclusion of further motivated parameters, and/or interactions between the parameters that are not immediately obvious to an observer.

In the following sections, I discuss approaches to analogical learning. At Section 1.3.2.3, I have already alluded to the connection between analogy and productivity; for my purposes, analogy is a formal morphological change that affects existing forms, whereas productivity in the broad sense can generate forms that previously did not exist. The processes that underlie a productive process, however, must result from generalizations that characterize existing forms; thus, the patterns that inhere in existing forms can both impact one another (analogy), and determine the outcomes for novel formations (productivity). My discussion here is principally devoted to explaining the operations of the Minimal Generalization Learner.

### 2.3.1 Analogical Learning Techniques

In the most general terms, theories of morphological learning and production look at morphologically complex forms as being produced either through a single general process, or by at least two entirely distinct processes. The latter theory, perhaps most thoroughly developed in work concerning the acquisition and modeling of English past tense production (especially Marcus et al. 1992, Prasada and Pinker 1993) is a *dual-mechanism model*, in which forms instantiated by the most general and widely available process are *all* assumed to be generated by a symbolic and abstract rule, while all other functionally overlapping forms are retrieved from memory by way of an associative pattern-matcher.[30] Under this dual-route

[29] Or, at least, the number of such errors must be small enough, and the overall quantitity of data large enough, that a model can be robust against such outright errors.

[30] This *dual-mechanism* view of morphological processing is to be distinguished from the *dual-route* model to be discussed at 3.4, which may assume a single-mechanism (i.e., rules or something else) as a means of

approach, the very fact that a form does not instantiate the most general rule *compels* it to be a memorized form; "irregularity" and non-systematicity is taken to be the cause of storage. In substance, this view is very similar to van Marle (1985)'s perspective on competing derivational processes (cf. 1.3.2.1 above): productive processes are a general case, and are productive to the extent that their potential domains of application are not partially blocked by "irregular" forms. While only the symbolic rule mechanism would be truly productive, the pattern-matcher might occasionally (probabilistically) suggest "analogical" forms.

In contrast, *single-route* models deny a distinction between symbolic rules and "analogical" pattern-matching. The productive part of the morphology must, then, be constituted by a multitude of rules, or something that simulates the appearence theoreof. The question is essentially whether an "analogical" pattern-matching process gives the appearance of rules, or whether highly constrained rules can produce outputs that have traditionally been attributed to analogy.

Approaches to the modeling of analogy that have been developed over the course of the past thirty years are all adequately described as systems of prediction: they attempt to specify some features of an output, given a particular set of input variables, through the same essential technique. Skousen (2002: 3) divides systems of prediction into two basic types: declarative (rule-based) systems and procedural systems; the former explicitly state the means by which a given input is converted into another output (and, Skousen observes, rules are probably necessary to synthesize and communicate about linguistic behavior), while the latter make no explicit statements about the regularities that hold in the data. Among such procedural approaches, Skousen further distinguishes between exemplar and non-exemplar based approaches. Non-exemplar approaches allow for and presume abstraction from the forms that appear to instantiate a process; exemplar models instead make predictions directly from interactions that obtain between exemplars, without any abstraction.[31]

Apart from the symbolic rule component of dual-mechanism models, none of the systems of prediction with which I am familiar are strictly declarative systems: they are all procedural models that either abstract away from the learning data (connectionist neural networks, Minimal Generalization Learning), or are exemplar-based models that deny abstraction (Analogical Modeling, Memory-Based Learning). Empirically, the performance of connectionist models (e.g., Rumelhart and McClelland 1986) is rife with problems; I refer the reader to criticisms in Pinker and Prince 1988 and Baayen 1989: Ch. 8. Empirical tests of Analogical Modeling (Skousen 1989, Skousen et al. 2002) are difficult to evaluate fairly, because they often exhibit a crucial design flaw. Albright and Hayes (2003: 122) rightly say "A model must fully specify its intended outputs;" if some full form is not selected from among all possible options, it is difficult to determine what, in fact, the predictions of the model are. The study of Eddington 2000, which employs only three class labels ("regular", "vowel-changing irregular", or "other irregular") to model English past tense production in the Analogical Modeling (Skousen 1989) framework, is thus clearly deficient in this regard. Other

morphological production.

[31] Immediately worrisome with respect to exemplar-based models, is how their products can be understood, granting the correctness of Kelly and Martin (1994: 116)'s claim: "The use of statistical regularities by animals also depends on abstracting away from individual tokens."

Analogical Modeling studies likewise appear to employ a compressed selection of class labels that may facilitate the classification task.

The two remaining models, the Memory-Based Learning model (MBL), as described in Daelemans and van den Bosch 2005, and Minimal Generalization Learning (MGL; Albright and Hayes 1999), show substantial overlap with one another; Keuleers (2008: 16) in fact says that "MGL can be seen as an implementational variant of MBL." Specifically, Keuleers means that "the pairwise comparison of verbs [or whatever kinds of formal inputs, RS]," which is the means by which MGL builds up the rules that it uses in evaluation, "is equivalent to a memory-based system, which compares the lexicon to each target form at run time" (Keuleers 2008: 151).

MBL is essentially a *k*-Nearest Neighbor (*k*-NN) classifier: it uses a distance metric[32] to assess the degree of similarity between a novel input and existing forms; the class label of the most frequent label in the number of neighbors evaluated is then assigned to the novel input. Thus, within a 3-NN model, if two neighbors belong to one class, and the third to a different class, the novel generalization is assigned to the class with the two members. At maximum *k* (i.e., the total number of data points), the method reduces to a type-frequency classifier; thus, in the case of English past tenses, all forms would be assigned to [-d] suffixation, since that is the most frequent output.[33] The requirement to generate fully specified outputs can be met by employing a sufficient number of class labels; Keuleers (2008: Ch. 4), for instance, uses $24$ distinct class labels to characterize every distinct surface pattern that appears in English past tenses.

I will describe the operations of MGL in more detail below. In empirical tests, MBL and MGL also produce similar outcomes; Keuleers (2008: Ch. 4), in assessing correlation of MBL and MGL to wug-test data from Albright and Hayes 2003, show overall indistinguishable outcomes.[34] Nevertheless, there are several theoretical and empirical reasons, at this time, to prefer MGL to MBL

First, Albright's research program (e.g., Albright 2010) has already shown effective integration of MGL with historical linguistic research – MGL is known to be able to reproduce attested historical changes. In principle, I see no reason why MBL would not be capable of similar predictions, but performance is this regard is as yet undemonstrated. More troubling is the fact that the actual implementations of MBL, while dividing segments into prosodic onset/nucleus/coda slots, do not employ the phonological features of the segments; hence, a sequence [bi:k] is evaluated as equally different from (much more similar) [pi:k] and (much

[32] In the Tilburg MBL implementation (`http://ilk.uvt.nl/timbl/`), the Manhattan distance metric is typically employed.

[33] Maximal *k* is equivalent to the Generalized Context Model (GCM; Nosofsky 1990), which employs an exponential decay function for weighting the relevance of the exemplars.

[34] The MBL simulations with the setting $k = 7$ performed by Keuleers on English past tense formation perform, on the whole, about as well as MGL, though the MGL still performs markedly better on Islands of Reliability for "irregular" forms than either the GCM or MBL. In Keuleer's implementation, Islands of Reliability for "regular" past tense forms do not fit well to the wug test data; in fact, she claims that the correlation between MGL performance and wug test responses on Islands of Reliability reported in Albright and Hayes 2003 is erroneous, caused by an error in scaling model results to wug-testing scores.

less similar) [gi:k] (though see Keuleers 2008: 84–5 for an attempt to partially circumvent this issue). Again, although in theory there is no reason why full sets of phonological features could not be employed in calculating distance in the MBL (which might aid performance), the absence of phonological features makes the behaviors of a trained model somewhat opaque.

The most serious defect, I believe, of wholly exemplar-based models, is the greediness of the memory. As Keuleers (2008: 158) says, "given the idea of an *exhaustive* (emphasis mine) storage of experiences, there is no reason why a memory-based system should try to generalize forms it already has in its memory." This view, I fear, makes the memory too powerful; speakers should not analogically remake forms to which they have been exposed; diachronically, we would see analogical changes occur, but they would necessarily be attributed to failures in transmission, i.e., a speaker did not receive exposure to the relevant data point at all. While some degree of bias in favor of forms to which a learner has been exposed is entirely reasonable (cf. Zuraw 2000's USE-LISTED constraint), the prediction of exhaustive storage is not. Furthermore, exhaustive storage would not predict the well-known U-shaped curve that occurs in the child's acquisition of morphology: if the child does not (or cannot) forget about exposure, or apply well-supported rules in the face of irregularities that do not (yet) have adequate support, we should never expect to encounter overregularization errors.

Moreover, as I will make clear in Chapter 3, I believe that a purely exemplar-based model of morphology is logically difficult to reconcile with the approach to morphological productivity that I adopt, and its corresponding psycholinguistic correlates. Namely, if we accept that certain processes are more productive because they are more psychologically active (i.e., the process itself is more readily accessible and usable because it is used more often), then the only means by which **the process itself** can be more active is for the process to be independent.

For these reasons, to operate with MGL as a predictive model of analogy is most sensible at this time. The underlying assumptions of the model are concordant with mainstream phonological theory and, I believe, with core observations on morphological acquisition. Most importantly, for present purposes, the literature on MGL demonstrates its relevance to the study of historical morphology.

### 2.3.2 Minimal Generalization Learning

MINIMAL GENERALIZATION LEARNING (MGL, first proposd Albright and Hayes 1999, and variously implemented in Albright 2002b, Albright 2002a, Albright and Hayes 2003, Albright 2005, Albright 2008a, Albright 2008b, Albright 2010) takes a perspective exactly opposite to that of Becker 1990: rather than trying to subsume morphological rules under analogy, analogy is explained as the extension of morphophonological rules to new forms. In Albright 2008b, the results of modeling analogical extensions in diphthongizing Spanish verbs (of the type [sen'tar] 'to feel' : ['sjenta] 'feels' or [kon'tar] 'to count' : ['kwenta] 'counts') using both MGL and the Generalized Context Model of Nosofsky 1990 (which Albright considers more "analogical" in its implementation, becuase it relies on "variegated similarity") give

preference to MGL. MGL operates with SPE-style (Chomsky and Halle 1968) rewrite rules, which have a "reliability" depending upon how many cases they successfully apply to, based on the structural description of the rule. Hence, if two rules have the same structural description, or the structural description of one would be a subset of the other, and the rule with the same or more general structural description has greater reliability than its identical or more specific counterpart, then a situation ripe for analogical replacements comes about. Because these rules have all have relative probabilities of application, and some may depend on very specific environments, one can determine with greater precision where an analogy might apply, and where not.

The MGL is certainly not unique, and is neither the earliest nor the most recent variety of a sort of automated learner for phonology and morphology. Indeed, Albright and Hayes (1999: 3) explicitly acknowledge their intellectual debt to early work on computational learning models. Other models, outlined in Daelemans and van den Bosch 2005 and McClure 2011, also describe results produced by their models in probabilistic terms that could be taken to describe the productivity of morphological patterns.

From a theoretical standpoint, the success of MGL has demonstrated that the problem of analogy can be made scalar and stochastic, just like the problem of productivity. Regardless of whether the symbolic rules with which MGL operates indeed have psychological reality, they provide a convenient, and empirically accurate means of discovering why certain analogies occur. Moreover, MGL offers yet another perspective that allows for one to group together rules and analogy, which then permits treating analogy and morphological productivity in a like fashion.

Insofar as analogy and productivity reflect essentially the same process, MGL may also be able to measure the productivity of morphological patterns; Albright and Hayes (2006) and Albright (2012) have explicitly observed this possibility. Unlike the broad categorial measures of productivity $\mathscr{P}$ and $\mathscr{P}*$, however, MGL aids the discovery of smaller and more local patterns that can be relevant to the analogical extension or productivity of morphological patterns.[35] This section describes the basic machinery of MGL, how it functions in terms of both intraparadigmatic analogy (discovering productive patterns in inflection) and interparadigmatic analogy (discovering productive patterns in derivation), and how frequency-adjusted reliability scores for rules ("confidence"; cf. Albright and Hayes 2003: 126–7) can serve as measures of productivity.

In the description of MGL that follows, I refer the reader generally to Albright and Hayes 1999 and Albright 2002b: Ch. 3 for further details. In its basic operation, MGL compares a possible or assumed mapping between two sets of surface forms (e.g., singular → plural in German nouns, present → past in English verbs) and attempts to compose morphophonological rules that successfully carry out that mapping. That is, the MGL examines pairwise (and only pairwise) mappings of forms in a dataset, and tries to extract the most specific possible rules that can generate the forms. The "tightest hypothesis that covers the range of relevant data" (Albright and Hayes 1999: 3) is a minimal generalization; at its most specific,

[35] Albright and Hayes (2006) addresses problems that arise from accidentially exceptionless generalizations that emerge from the data, which generate local rules that make incorrect predictions for novel forms.

the mapping X → Y, where X and Y are two forms that stand in a pardigmatic or derivational relationship to one another, is equivalent to the memorization of just that pairwise mapping. The learner, however, attempts to discover rules that have more general applicability, having not only a structural change, but also a structural description; often, the learner arrives to very general default rules that apply the structural change in any environment (e.g., the rule Ø → d / X_# [where X is any number of segments] is learned for the mapping of present → past in English verbs).

When attempting to discover the environment for a rule, the learner employs edge-in alignment to uncover similarity between a pair of mappings. For instance, from the mappings kɪs → kɪst and mɪs → mɪst, the respective minimal rules Ø → t / kɪs_# and Ø → t / mɪs_# emerges. From those two mimimal rules, a more general rule Ø → t / Xɪs_# is generalizable. Note here that not only identically matching segments, but shared features can be learned as well. As the learner incorporates and generalizes across more and more minimal rules that share the same structural change Ø → t, it may well arrive to the general Ø → t / X_#, but it will still retain the more specific rules that apply to at least two inputs.[36] At this phase, the reliability and confidence of rules come into play. Reliability is a simple ratio defined as the number of HITS (correct outputs predicted by a rule) divided by the SCOPE of the rule (i.e., the number inputs captured in its environment). Confidence is an adjustment of reliability, following a procedure described in Mikheev 1997 using lower confidence limit statistics (cf. Albright 2002b: 40, fn. 5), according to the size of the scope. In effect, two rules that have the same raw reliability score can have very different confidence levels, depending upon their scope; using a 75% confidence interval,[37] a rule with reliability of $.5$ but a scope of only $2$ will have a confidence of just $.146$, whereas a rule with the same reliability but a scope of 1000 will have a confidence of $.489$. In principle, the rule with the highest confidence score that has scope over an input is the rule that a learner ought to select. "The default mapping for any given context is the one with best Confidence" (Albright and Hayes 1999: 5).

Where the rule with greatest confidence generates an output that is different from the output in the original data, an analogical change is predicted to occur. In effect, an analogical change can simply be the extension of a morphophonological rule to a form where it did not previously apply. Within inflectional paradigms, Albright (2002a, 2006, 2010) has adduced evidence from Latin, Yiddish, Lakhota, German, and Korean that single paradigm members function as a BASE from which the other paradigm members are generated. In principle, the most informative paradigm member (i.e., the member that can most often correctly predict the expected forms of other paradigm members) should function as the base; this informative base usually shows few or no phonological neutralizations. When no paradigm member is able to predict all the expected forms of the paradigm, the situation is ripe for analogical changes.

The confidence of probabilistic context-sensitive rules as applied to derivational pat-

[36] In this way, rules supported by just single forms, which would mimic the behavior of a 1-NN classifier using the MGL's similarity metric, are wholly deprecated.

[37] Albright and Hayes (1999: 26) make "a very rough estimate, based on Wug testing data, that the appropriate confidence limit [to be applied to linguistic data] is about 75%."

terns could similarly be predictive of the productivity of that pattern. The notion that bases exist in derivational morphology is uncontroversial; morphologically complex words are presumed to be (at least etymologically) derivative of corresponding morphological simplexes. Indeed, "word-families" built on the same root may give the appearance of quasi-paradigmatic relations between derivationally related forms (cf. Aronoff 1976: 40-5 on English *-ous*, *-ity*, *-ness*), and the Indo-European "Caland" system[38] seems to presuppose this possibility as well. In the case of apparently competing derivational processes, the "Island of Reliability" (see Albright 2002a) effects that MGL captures would successfully model the domain-restricted productivity of certain processes (e.g., the productivity of English *-ity* to adjectives in *-able/-ible*, instead of the more generally productive *-ness*).

Since reliability and confidence in MGL rest essentially upon the type frequency of a rule, they must capture some different facet of productivity than what $\mathscr{P}$ and $\mathscr{P}*$, which take account of token frequencies within types, measure. MGL might then, in some cases, predict the productivity of a process simply based on high type frequency. However, type frequency ($V$) in Baayen's research regularly correlates with $\mathscr{P}*$, so confidence and $\mathscr{P}*$ might prove to make similar predictions. On the other hand, in dealing with competing derivational processes, the number of actual hits for a rule following its environment will be well below unity. To return to the example of English abstracts in *-ness* and *-ity*, the general rules Ø→ nɛs / $X_{adj\_}$# and Ø→ ɪti / $X_{adj\_}$# will both have a reliability (and confidence) of less than 1, but the reliability (and confidence) of Ø→ ɪti / $X_{adj\_}$# will surely be less than Ø→ nɛs / $X_{adj\_}$# – and presumably the more minimal rule Ø→ ɪti / $Xəbl_{adj\_}$# will have a confidence exceeding general Ø→ nɛs / $X_{adj\_}$#. From this point of view, MGL could render results that resemble $\mathscr{P}$.

Comparison between productivity as measured by confidence and as measured by $\mathscr{P}$, $\mathscr{I}$, and $\mathscr{P}*$ is thus a topic open to exploration. In general, I expect that the measures will make roughly similar predictions, but work with probabilistic context-senstive rules may aid in the discovery of some more subtle patterns, or reveal some fine details that motivate the productivity of a larger process, as measured by $\mathscr{P}$ and $\mathscr{P}*$. Taken together, Baayen's productivity measures and MGL can provide an effective toolkit for the exploration of productivity, as defined under 1.2.1.

[38] See Caland 1892 and Nussbaum 1976. The "Caland" system describes sets of derivationally related forms where a morphologically simplex base that could serve all of those forms is often unattested.

# CHAPTER 3

# The Psycholinguistic Bases of Productivity

Reasoning about the problem of productivity in Chapter 1 concluded decisively that a probabilistic interpretation of "productivity" is best able to render reasonable results that admit of comparison cross-categorically. More generally, I argued, following de Saussure, that "productivity" and "analogy" are really two sides of the same coin, and so both their theoretical treatment and specific analysis should not be treated independently.[1] In Chapter 2, I introduced the research of Harald Baayen, and the statistical corpus-based techniques that he has developed as a measure of productivity; I also discussed the Minimal Generalization Learner of Albright & Hayes, which provides a means for perspicuously identifying smaller, yet potentially extensible, patterns.

However, the crucial question that remains is: what psychological reality do any of the tools and techniques discussed in Chapter 2 actually have? All of the measures and methods discussed in the preceding chapters fundamentally presume that human linguistic competence both tracks frequency of linguistic material and is sensitive to statistical distributions in language. Of more direct import to our studies is the question: do we have good reason to believe that those measures and methods provide reasonable and interesting approximations of linguistic phenomena?

The short answer to this question is "yes". Frequency effects are amply attested in the literature on morphological processing specifically (see the papers in Baayen and Schreuder 2003), and the psychological literature more generally attests to robust sensitivities to statistical distributions among humans both domain-generally (cf. Hasher and Zacks 1984) and in language in particular (cf. Saffran 2003 and papers in Divjak and Gries 2012).

Baayen (1989: Ch. 7) summarized much work on morphological processing up to that date, and undertook to interpret how his proposed measures of productivity correspond to results reported in the psycholinguistic literature. This chapter now aims in part to update Baayen's literature review.[2] In the main, I refer the reader here back to Baayen's 1989 summary for older literature on morphological processing.

Discussion in this chapter thus assumes the following plan:

[1] I do, however, prefer the term "productivity" to cover both. Forms that "replace nothing" instantiate raw productivity; apparent replacements or alterations of existing forms result from the application of more productive processes. In the case of intraparadigmatic leveling, the issue concerns both selecting the best base and the most reliable inflectional rules.

[2] At the same time, it is worth noting in advance that many psycholinguistic studies on morphological processing during the past twenty-five years often make explicit reference to Baayen's work, or indeed, include Baayen as a co-author.

- I first review some summaries of the psychological literature on learning in both animals and humans that offer strong and systematic evidence for frequency-matching effects and statistical learning. These results indicate that statistically based effects in any aspect of human cognition, including language, are entirely expected.
- I then examine the literature on frequency effects on morphological processing and production specifically. Effects of token frequency and type frequency, at the level of individual lexemes and in relations between lexemes, are all considered.
- Independent of but closely related to these issues of morphological processing are the connections of phonological and semantic transparency. I review the possible fashions in which frequency and transparency are mutually facilitatory or inhibitory in language processing.
- The effects surveyed with respect to morphology are then interpreted in the light of the Morphological Race Model (MRM; Frauenfelder and Schreuder 1992, Schreuder and Baayen 1995). The MRM provides a plausible framework in which to assess the psychological status of lexemes and morphological processes based upon the frequencies and morphological relationships that can be extracted from corpus languages.
- Finally, I review the specific attempts to furnish independent psycholinguistic validation for the predictions of both the corpus-based productivity measures and the Minimal Generalization Learner.

## 3.1 Frequency, Probability, and Learning: Domain-General and in Language

The systematic study of both animal behavior and human psychology during the course of the past century offers powerful evidence that "frequency knowledge plays a critical role in many implicit cognitive functions" (Zacks and Hasher 2002: 34). In general, the literature reveals in both animals and humans a capacity for numerosity (i.e., awareness of and ability to cognitively process operations involving number) and intuitive frequency-matching behaviors. The very evidence for numerosity competence and frequency matching together constitute evidence for frequency tracking: without the ability to automatically attend to how often some event occurs in an environment (i.e., calculate *rate*, which requires both numerosity and temporal tracking), the existence of frequency-matching behaviors would be inexplicable. The extremely broad perspective that I give here is intended to establish the fundamental and basic role that frequency information plays for organisms, so that, when I turn to the role of frequency in morphological processing, the fact that the results obtained there imply a rich and fairly detailed representation of word and morpheme frequencies should not come as a surprise, but instead appear to be entirely in line with animal psychology as a whole.

In the first place, it is interesting to note that apparent frequency-tracking behaviors are not unique to humans whatsoever; indeed, frequency-sensitivity appears to be a crucial

evolutionary development, one which may partially underlie and certainly interacts with many other seemingly basic cognitive functions, such as the perception of time and space. See Gallistel 1990 for detailed discussion of animal learning, and Shettleworth 1998: 363–77 on animal frequency sensitivity specifically.

Indeed, in research on animal behavior, frequency matching has been replicated sufficiently often and with such regularity that it has assumed the status of a law, namely "Herrnstein's matching law", after the original work of Herrnstein 1961 on pigeon foraging behavior in an experimental setting (cf. Gallistel 1990: 361–3 for the status as "law"). Particularly clear evidence for an ability to calculate rates and thus determine the probabilities of rewards comes from experimental work with rats (Sutherland and Mackintosh 1971: 406 ff., summarized in Gallistel 1990: 351–2 and Kelly and Martin 1994: 109–110) and in a naturalistic setting with mallards (Harper 1982, summarized in Gallistel 1990: 356–9). In both cases, one can observe that individual rats or mallards possess a precise awareness of the rate of return at potential food sources, and that the mallards, as a flock, precisely track the extent to which food sources are exploited by the other mallards.

The experimental design involving rats is straightforward: the rat is placed at the bottom of the longest end of a T-shaped maze, and, depending upon which among the left or right arm of the T is "correct" on a given trial that the rat enters, the rat will receive a piece of food. If the probabilities of the "correct" arm are different, e.g., in 75% of trials the right arm is "correct" while in 25% of trials the left arm is "correct", provided that the rat is able to know after the fact which side was "correct", the rat's choice of arm over a series of trials approaches the distribution of arms. That is, on 75% of trials, the rat will enter the right arm, but enter the left arm on 25% of trials.[3] Curiously, this behavior does not maximize the rat's long-run probability of obtaining a piece of food. The optimal behavior, in that sense, would be to always select the right arm, since the rat would then receive food in 75% of the cases; instead, by alternatively selecting the left arm in 25% of the trials, the rat reduces its overall rate of reward to 62.5% (say, in 100 trials, $0.75 * 75 + .25 * 25 = 62.5$).

The basis for the rat's learned behavior of matching the probability of return for any given trial, instead of the long-run optimal possible return, becomes clear from how animals treat food sources in aggregate, i.e., when multiple individuals are competing for all food available in the environment. For instance, the flock of thirty-three wild mallards studied by Harper (1982) tended to divide themselves into appropriate ratios in response to both the rate and weight of bread morsels provided by experimenters at different points along the shore of a lake. Within one to two minutes upon the commencement of bread-throwing, with pieces of equal weight, the flock would achieve the appropriate ratio, i.e., when the rates of throwing were equal, the mallards divided themselves equally, or when an experimenter threw bread at twice the rate (thus doubling the profitability of that food source), the mallards divided themselves approximately 2:1. Similarly, when the morsels being thrown by one experimenter were twice as large (but at equal rates), the mallards (after initially distributing themselves according to rate), assumed a 2:1 ratio of distribution within five to six minutes.

[3] Gallistel (1990: 352) also anecdotally notes that undergraduate students, able to see which of the arms in the maze was correct after the rat had chosen an arm, when asked to predict the behavior of the rat in a series of trials, exhibit the same probabilistic behavior as the rat: a 75% to 25% distribution.

Thus, whenever differential quantities of food are available at multiple local sources, the animals seem to act in concert to exploit maximally these available resources; no animal acts as though the others do not exist, and tries to monopolize only the more profitable source. The crucial point, regardless, is that the mallards must be able to discriminate temporal intervals, number, and magnitude, with respect to both the food sources and all of the competitors for the food, and accordingly match frequency distributions such that the probability of obtaining food is maximized by all.

Among humans, controlled experimentation has repeatedly confirmed that individuals do, with a good degree of accuracy (i.e., close correlation between real values and responses from subjects), match frequencies; the evidence accumulated in the psychological literature is virtually unequivocal. See Hasher and Zacks 1984 for a good and thorough general summary. These results have been obtained mainly from three experimental paradigms: *frequency judgments*, which ask subjects for a direct estimate of an item's frequency, *forced choice* studies, which ask subjects to decide which among one or more items is most frequent, and *ranking* tasks, which ask subjects to list items in terms of frequency. Performance in such tasks appears to be unaffected by experimental instructions, i.e., regardless of whether or not subjects are told that their memory of stimuli frequency will be assessed; this issue is specifically considered in Burnett and Stevenson 1979.

One early study, Attneave 1953, elicited frequency judgments for each of the 26 letters used in writing English from adult speakers of English, as frequencies per 1000 letters; the relative frequency ranks given were fairly accurate, and Attneave determined the correlation between the true frequencies of letters in English words and the frequencies thereof estimated by subjects to be $r = 0.79$.[4] This healthy correlation demonstrates awareness of frequency for some objects present in the environment that are surely not deliberately and consciously tracked. Hintzman (1969) further clarified, in line with Attneave 1953, that relative frequency awareness is strong, though absolute frequency reporting is subject to distortion – apparent frequency reported by subjects resembles a logarithmic function of actual frequency (very frequent items are underestimated, while infrequent items are slightly overestimated; Hintzman 1969: 141). In Hintzman's study, subjects were asked to give frequency judgments or make a forced choice between two items (in this study, words), to which they had been exposed during a training period. Similarly, a set of three experiments using forced choice tests in Zacks et al. 1982, in which subjects were exposed to lists of 90 English words occurring between 1 and 7 times, found that subjects correctly selected the more frequent item from the presentation period in $\sim 80\%$ of cases. Over a longer training period of two weeks, Hasher et al. (1977) found that the frequency with which subjects were exposed to statements on a range of topics (including politics, sports, the arts, medicine, geography), regardless of whether the statements were in fact true or false, increased the likelihood that the subject would judge a statement to be true.

This small sample of studies convergently points towards an inherent capacity for the tracking and relatively fine discrimination of frequency in humans. The fact that this fre-

[4]The participants tended to underestimate the total number of letters, thereby probably weakening the correlation; the mean of the sums of the elicited frequencies was only $896$, rather than the expected $1000$.

quency knowledge can be metalinguistic and experiment-external (letter frequencies), language-related and experiment-internal (word lists), or involve real-world propositional knowledge, is indicative of a domain-general skill running throughout cognition. Combined with the evidence from animal behavior, we have powerful evidence that frequency sensitivity is an innate capacity, employed as an organizing principle in learning and memory, and can be used to guide and shape future behavior. Note, moreover, that certain tasks, such as estimating the frequency of letters in a 1000-letter sample sample, given the population of letters to which an individual has been exposed, are indicative of the ability to make probabilistic inferences based upon the frequency data that the individual has processed.[5]

### 3.1.1 Frequency Sensitivity: Why and How?

Given the above-summarized facts, a sensible question would be: why would creatures on earth have developed such pervasive frequency-tracking behaviors? What evolutionary advantage could (subconscious) statistical awareness confer? If frequency-sensitivity is a commonplace feature, one might wonder: what adaptive function does it serve? Most importantly, what are the necessary features of a model of cognition that can account for these behaviors? Kelly and Martin 1994: 107 offer a sensible general assessment: "animals must often make rapid decisions while minimizing the likelihood and severity of errors. These considerations lead to the following conclusion: animals that have the capacity to detect probabilistic patterns and exploit them will have an advantage over those that do not. Furthermore, confidence in the solution to a problem should increase if multiple cues converge on that solution." This view makes frequency matching part and parcel of evolutionary psychology. For both humans and many animals, registering such information seems to be automatic and pervasive, and yet humans and animals do not seem to have a conscious awareness of the fact that they register this information. Nevertheless, there is no real consensus on *how*, precisely, frequency information is cognitively stored; Hasher and Zacks (1984: 1379–80) discuss both a trace theory (Hintzman 1976) and a counting mechanism (Underwood 1983). This issue is not, however, crucial to present considerations.

A clear adaptive advantage emerges in the consideration, again, of foraging behavior. Animal foraging behavior approximately adheres to a *Marginal Value Theorem* (Charnov 1976). In order for the animal's foraging behavior to exploit and make sense of marginal value, it must, at minimum, have a conception of the average rate of return within the zone that the animal inhabits, and how efficiently it is able to forage from moment to moment. Shettleworth (1998: 374) summarizes:

> To maximize its overall rate of energy intake, a forager should leave a patch when the rate of energy gain in that patch falls to the average rate in the habitat. Before this time, the forager is by definition doing better than it can elsewhere.

[5] Despite the ample evidence that humans are good probabilistic learners, humans nevertheless often fall victim to various fallacies in probabilistic reasoning. See, though, Chater et al. 2006 for how the likely probabilistic operations of cognition are to be reconciled with such fallacious reasoning. Note also that direct evidence of probabilistic reasoning has recently been obtained not only for human infants, but also for great apes (Rakoczy et al. 2014).

> Afterward, it could do better on average by leaving. In order to behave in this way, a forager has to keep track of its current rate of intake. If prey come in discrete similar-sized units like leatherjackets[6] in a field, this means accumulating information about times between prey captures. The forager also needs information about average intake rate in the rest of the habitat.

Of course, the fact that animals track rates with respect to some natural phenomenon that directly impacts survival (such as the richness of a food source) does not entail that they track the rates and frequencies of everything in their environments. However, experimental associative learning indicates that animals (such as the rates of return described above) can and do attend to the frequencies of any phenomenon that pertains to rewards or punishments. At the outset of a course of learning, an animal cannot know that the occurrence of some event will necessarily prove to be relevant or irrelevant; nevertheless, the learned patterns of behavior match the rates of occurrence from the outset – there is not a point at which the animal begins to attend to the phenomenon, having realized its potential relevance. The conclusion is that frequencies are automatically tracked with minimal effort on the part of processing or memory.

This conclusion is built in to the Automatic and Effortful framework developed by Hasher and Zacks (Hasher and Zacks 1979, Hasher and Zacks 1984, Zacks and Hasher 2002). One important general point established by Hasher and Zacks is that no significant differences exist in performance involving the memory of temporal, spatial, and frequency characteristics of events between persons of all ages and abilities (i.e., it is developmentally invariant), under both stressful and calm conditions. People under a very broad range of circumstances reliably and unintentionally encode information about the relative frequencies of events.[7] Hence, Hasher and Zacks conclude that the registration of information pertaining to frequency in memory must be automatic and require extremely minimal cognitive effort – "automatic operations function at a constant level under all circumstances." (Hasher and Zacks 1979: 356). Moreover, performance on tasks involving capacities that Hasher and Zacks consider to be automatic does not improve with deliberate training, and is neither improved nor hindered by experimental instructions that could direct or distract subjects. In certain "core domains" (e.g., processing of spatial, temporal, and frequency information, and perhaps the use of language itself), it thus seems that "human cognition approaches an optimal level of performance" (Chater et al. 2006: 289).

### 3.1.2 Frequency Effects and Statistical Learning in Language

With the facts of frequency sensitivity and frequency matching across general domains now in hand, I would like to consider similar evidence that pertains to the domain of language specifically. Since the relevance of frequency to morphology and morphological processing will be treated in greater detail under 3.2, I will cite here evidence from other areas of

[6] A common term for the larva of a large crane fly (genus *Tipula*).

[7] Indeed, Hasher and Zacks anecdotally report that experimental subjects are often, at first, resistant to perform frequency estimation tasks because they do not believe themselves to possess the relevant knowledge.

linguistic research. Ellis 2012 (especially pp. 11–5) is an up-to-date overview of frequency effects in linguistic research from a Usage-Based perspective. Jurafsky 2003 also offers helpful discussion with respect to psycholinguistics specifically, and Saffran (2003) very briefly summarizes some experimental literature and marshals arguments that support the conclusion that many basic and essential language acquisition tasks, such as word segmentation, are statistically driven. Specific papers in Bod et al. 2003 are also worth consulting for more detailed discussion on the role of frequency and statistics in all major subfields of linguistics.

The centrality of frequency to the operation and construction of grammars is evident from some empirical observations, and a number of experimental results. First, virtually all languages exhibit some range of processes that apply variably or "optionally"; for instance, /s/ in syllable codas throughout many varieties of Spanish undergoes lenition to [h] or Ø at varying rates and under various conditions (cf. Bybee 2007: 220–4 and references therein). Optional applications appear not only across the population of speakers (i.e., variability is not a phenomenon of the *parole* that is emergent from categorically different *langues*), but within single speakers. At minimum, learners must attend to relevant distributional cues in order to determine the likelihood of a process' application. It is for this reason that Ellis (2012: 9) states:

> Frequency is a key determinant of acquisition because "rules" of language, at all level of analysis from phonology, through syntax, to discourse, are structural regularities which emerge from learners' lifetime unconscious analysis of the distributional characteristics of the language input.

Furthermore, experimental results employing both artificial language learning and "wug" testing[8] reveal both a ready ability to make use of and acquire patterns according to the rate that they occur in the data, as well as apply patterns following their rate of occurrence in some real language that a speaker knows. Once again, these facts would be unexpected and difficult to explain if humans did not constantly attend to and automatically record frequency data in their environments. I now review some evidence of this sort from phonetics/phonology and syntax before turning to the specific role of frequency in morphological processing.

#### 3.1.2.1 Phonetics and Phonology

A substantial number of studies on the acquisition of phonetic categories generally, language-specific phonological categories, and the problem of (phonological) word segmentation considers the role of frequencies overall and co-occurrence frequencies of units as crucial factors. Already between the ages of approximately 8 and 11 months, infants exhibit decided phonotactic preferences that accord with the language being acquired (see the literature cited in Saffran et al. 1996: 606–7 and Hayes 2004: 161), as well as the ability to segment units that roughly correspond to phonological words (cf. Saffran et al. 1996: 608).

[8] Read under 3.2 for the description of a "wug" test.

With respect to the discrimination of phonetic categories in a language, Maye et al. (2002: B103) rightly observe that the tokens intended as the members of one category (e.g., [t]) will be more similar to one another, on average, than tokens intended as members of another category (e.g., [d]), that differ in a particular fashion (in this case, in terms of voice onset time (VOT)). Given that voice onset time is indeed a useful discriminant of different categories in the language, it can be used, statistically speaking, as a way of organizing the two categories into a bimodal distribution – there will be relatively few phonetic tokens that lie between the prototypical zones of the categories along the VOT dimension. As expected, Maye et al. (2002) then demonstrated (through use of a preferential looking procedure; cf. Aslin 1995) that both 6- and 8-month old infants show discriminatory behavior (i.e., perception of a contrast) when trained on stimuli that exhibit a bimodal distribution, but not when trained on stimuli with a unimodal distribution.

Saffran et al. (1996), meanwhile, explored the hypothesis that transitional probabilities between syllables might serve as effective cues to word boundaries, through an artificial language learning paradigm with university students. Using a set of 12 CV syllables made up of four distinct consonants and three distinct vowels to compose six trisyllabic words (e.g., [pidabu], [dutaba]), designed in such a way that all word-internal syllable transitions would be more probable (at least $0.31$) than any possible syllable transitions between words (at most $0.2$). In a training period, experimental subjects were exposed to three hundred tokens each of randomly concatenated words in synthesized speech over twenty-one minutes. During a testing phase, subjects were presented with $36$ tokens in total of the designed words, trisyllabic sequences that could not have arisen through mis-segmentation of the speech stream, and trisyllabic sequences that could have been perceived through mis-segmentation. Subjects were asked to identify whether the item did or did not sound like a word heard during the training phase; performance on all testing conditions showed overall performance better than chance ($p < 0.01$). Analysis of performance on individual items revealed that performance on the three words with the higher transitional probabilities was significantly better than on the three words with lower transitional probabilities. These results strongly favor the interpretation that rarer transitional probabilities serve as effective indicators of phonological word boundaries, and that humans happily seize upon such evidence.[9]

Two clear instances of frequency matching behavior emerge from studies involving "wug" items with native speakers on the distribution of phonological entities in the lexicon and the application of phonological processes. Within the Dutch lexicon, for instance, the frequency distributions of underlying stem final voiced and voiceless segments are appreciably different: in data drawn from the CELEX corpus (as reported in Ernestus and Baayen 2003: 9), only $9\%$ of stem-final labial stops are underlyingly voiced, while $97\%$ of velar fricatives are underlyingly voiced. Since Dutch systematically devoices underlying voiced obstruents word-finally, in principle, whether a stem ends in a voiced or voiceless segment can only be discerned from forms with a suffix (e.g., noun plural [-ən] or past tense [-tə]/[-də]). Ernestus and Baayen (2003) presented speakers of Dutch with "wug" verbs embedded in small carrier phrases (e.g., [ɪk dɑup] 'I *daup*'), such that the underlying nature of the obstruent would

[9] Similar evidence, we will see at 3.4, may be used in the detection of morphological structure.

be unknown, and asked the participants to generate a corresponding past tense form with the suffix [-tə]/[-də], to diagnose whether the participant regarded the stem-final obstruent as voiced or voiceless. Logically, since form like [dɑup] could reflect either /dɑup-/ or /dɑub-/, one might expect expect speakers, in absence of clear evidence, to randomly assign either /dɑup-/ or /dɑub-/ as the UR for [dɑup]. Instead, the participant's behavior closely matches the actual distribution of stem-final obstruent voicing quality in the lexicon: thus, Dutch speakers are much more likely to interpret [dɑup] as /dɑup-/, but [dɑux] as /dɑuɣ-/.[10] These results appear clearly in histograms in Ernestus and Baayen 2003: 17.

In a similar vein, Hayes et al. 2009[11] examined the factors controlling the application of vowel harmony in the Hungarian dative suffix /-nɔk/, which shows the allomorphs [-nɛk] and [-nɔk], under front and back harmony effects, respectively. While back vowels and front rounded vowels in a stem almost always obligatorily trigger the appropriate harmony, zones of variation exist. The authors were able to accurately model the distribution of these two affixes in a corpus using a logistic regression model that incorporated both phonetically natural (e.g., AGREE(back, local), AGREE(front, nonlocal)) and unnatural constraints (e.g., USEFRONT/ bilabial__), which reflect reliable patterns in the corpus data (essentially equivalent to Islands of Reliability; cf. 2.3.2 above). A forced-choice "wug"-test found that speakers of Hungarian approximate the frequency distributions of the respective allomorphs in novel forms, to an extent that would be difficult without awareness of and sensitivity to phonetically unnatural, but reliable, patterns that hold in the Hungarian lexicon.[12]

#### 3.1.2.2 Syntax

In parallel to the discrimination of phonological contrasts achieved by infants, distributional cues show profound relevance for the acquisition of syntactic structure as well. Gómez and Gerken 1999, based on a series of artificial language learning experiments with one-year-old infants using head turn preference paradigms, present compelling evidence that statistical learning mechanisms are employed to develop abstract word classes out of the dependencies that hold in the learning data.[13] Likewise, Saffran 2002 demonstrated that the syntax of an artificial language could be learned by both adults and children substantially better if it contained predictive dependencies (e.g., an article indicates a noun to follow), than if there are not predictive dependencies.[14] Both of these studies indicate that the co-occurrence

[10] Besides the statistical distribution of stem-final voiced segments depending upon place and manner of articulation, also the nucleus of the stem-final syllable, and whether a vowel, sonorant, or obstruent immediately precedes the stem-final obstruent are relevant factors; these distributions also showed frequency matching behaviors in the "wug"-test.

[11] Other good discussion of frequency-matching phenomena in phonology appears in this article, pp. 825–6.

[12] Perhaps more importantly, the results of the "wug"-test can be captured more precisely, Hayes et al. (2009: 845–55) show, by incorporating a bias against the phonetically unnatural constraints – speakers learn constraints like USEFRONT/ bilabial__ because they reflect genuine patterns in the learning data, but tend to deprecate them.

[13] The results of these studies are further situated in the context of experimental work with infants in Gómez and Gerken 2000.

[14] In a sense, this result is totally unsurprising, since all languages do, of course, contain such dependencies in their syntax (even discourse-configurational languages like Ancient Greek have syntactic restrictions).

likelihoods of certain words, or abstractly formed categories of words, are useful as syntactic learning cues.

Analogous to the work on frequency matching in phonology, Bresnan and Ford (2010) experimentally tested the reactions of Australian and American speakers of English to dative constructions (i.e., prepositional datives with *to* versus double object constructions) in the light of the frequencies of different constructions in corpus data. Specifically, a logistic regression model using features such as the animacy and pronominality of themes and recipients that was trained on corpus data makes predictions that parallel experimental results. Different groups of subjects performed a rating task, in which they indicated the "relative naturalness" of parallel sentences, one version with a prepositional dative, the other with a double object construction, and a continuous lexical decision task. In the rating task, sentence types that matched and were more frequent in the corpus data were rated more highly, and likewise elicited faster response times in the lexical decision task. These results strongly indicate that speakers are sensitive to probabilities of dative constructions in a way that parallels their actual occurrences in corpus data.

#### 3.1.2.3 Local Summary

We now have encountered rich evidence that frequency distributions in learning data must be exploited by humans in the process of grammar construction; in effect, those frequency distributions are the determinants (perhaps along with some innate biases) of how, precisely, to weight the factors of their grammars. The assumption that frequency distributions make for a powerful influence on future linguistic behavior emerges clearly from experimental work with both infants, thereby strengthening the claim of its relevance in the process of acquisition, and adults, thereby demonstrating that such knowledge is incorporated into relatively stable grammars. Furthermore, the close matches in behavior between stochastic models built upon corpus data in the domains of both phonology and syntax with native speaker judgments is very reassuring. Such results confirm that distributions in corpora, when subjected to linguistic analysis, prove themselves to be good approximations of the competencies of actual speakers. For the historical linguist who must subsist on corpus data alone, these implications are all the more felicitous: models based on corpus data, using factors whose linguistic relevance has been confirmed experimentally, can be trusted as a means of capturing variable data.

---

Therefore, whether the ready acquisition of a syntax with predictive dependencies follows from the reliability of the conditional probabilities, or an innate receptiveness to such structuring of the input is hard to say. However, the same bias towards learning with predictive dependencies appears in learning of non-linguistic materials, Saffran (2002: 186–90) shows; therefore, this use of distributional predictive dependencies may be a domain-general learning skill that humans possess.

## 3.2 Frequency Effects in Morphological Processing and Production

That frequency, in some sense, affects lexical processing is perhaps one of the best established results of psycholinguistic research (see already Howes and Solomon 1951). Generally speaking, the question under consideration is the extent to which word frequencies as attested in corpora, or the behaviors predicted by a computational model, significantly correspond to results from psycholinguistic experimentation. With respect to frequency, there are at least three distinct types of token frequency that can be considered: the absolute token frequency of a lexical item, the relative token frequency of an item to some other item (specifically, its apparent derivational base), and the token frequency of an item relative to some "family" of words (i.e., the frequency of forms derived to the same root, or derived through the same process). Type frequency, on the other hand, is only comparable at a broader level: how frequent is some process within the system as a whole?

Much of the literature on morphological processing specifically is aimed towards answering the question of what the underlying architecture involved in morphological production and processing may be (for a summary of six different views, ranging from a fully lexicalist [no analysis, only storage] to fully analytical [any item with any seeming components is always decomposed], see Diependaele et al. 2012). By and large, this research attempts to establish what relevant morphological effects influence the speed of word recognition (i.e., what factors facilitate or complicate word parsing and access); moreover, the research must attempt to disentangle other potentially relevant factors (e.g., orthographic, semantic) from the morphological factors. At present, the most useful reference on the field of morphological processing remains the edited volume of Baayen and Schreuder 2003, and I will draw heavily on the papers contained therein.

Throughout this survey, I will report on the concrete experimental results (e.g., the correlation between reaction time and some frequency), and the author's theoretical interpretation. At Section 3.3, however, I will attempt to unite all of evidence examined here within the framework of the Morphological Race Model. This model, insofar as it makes predictions about the fashion in which a given lexical item is likely to be treated in processing and production, can translate directly into an account of morphological productivity itself.

Before entering into discussion on particular frequency effects, it will be helpful to be familiar with several common types of experimental paradigms that are regularly employed in work on morphological processing, as well as standard views on how to interpret the results that emerge from those different types of paradigms. A variety of experimental paradigms exist in order to test lexical access. Very common are *lexical decision tasks*, which ask a subject to determine whether a lexical item is or is not a word of his/her language; presentation may be either visual or auditory in modality. In such tasks, longer response times are taken to indicate longer processing time. The tracking of eye fixation or eye movement is also employed, in which longer fixation on a word is interpreted as evidence for longer processing time.

Tasks that may relate more directly to type frequency, and more directly capture productivity, are the rating of nonce words or "wug" tests (Berko 1958), in which subjects are led to

inflect or derive novel surface forms from invented roots or stems. Rating tasks may simply present a list of words (together or in isolation), and ask the subject to indicate how "good", "bad", or "likely to use" a word seems to the subject. "Wug" tasks typically present some morphological base in an utterance (either the base word alone, if possible, or in an inflected form that gives no information about class membership) and ask subjects to produce a form within another sentence. For example, in the "wug" test given to speakers of Italian in Albright 2002a, subjects were presented with verbal stems inflected in the first person singular present (ambiguous as to class membership of the stem), and asked to produce an infinitive form (to determine the subjects' interpretation of the most likely class membership), embedded in a carrier phrase.

One general caveat concerning these psycholinguistic results is worth bearing in mind: they have been obtained almost exclusively from work on present-day Indo-European languages of Europe (especially English, Dutch, German, and Italian), though I am acquainted with some fewer studies on Finnish, Hebrew, and Japanese.[15] However, I have not seen any reason to believe that the results from these studies do not hold general typological validity (though see concerns in Marslen-Wilson et al. 1994: 4). Occasionally, one encounters claims to the effect that morphological processing plays a qualitatively different role in languages of different sorts in lexical access: Frost et al. (2005) claimed that lexical access in present-day Indo-European languages is primarily orthographically driven, but morphologically driven in Semitic languages; the conflicting study of Perea et al. 2014 on similar Arabic data instead suggests that these differences are merely quantitative. Thus, while the overall typological generalizability of all the experimental results to be discussed here is not wholly certain, the default position is that the results indeed have relevance cross-linguistically.

### 3.2.1 Lexical Token Frequency Effects

The impact of token frequency upon processing speed for roughly surface-level units in a language is among the best established results in psycholinguistic work. Gries (2008: 428) remarks that "it is well-known that (logs of) observed frequencies are good proxies towards the familiarity of words given the strong correlations of frequencies with processing speed."[16] However, we must distinguish between *whole-word frequency*, the token count of a specific surface phonological representation (distinguishing homophones), and *lemma frequency*, the token count of a stem minus any distinct inflected forms of the lexeme. The effect of

[15] On "wug" tasks in Japanese in particular, see Vance 1987: Ch. 12 and Kawahara and Shin-ichiro Sano 2014.

[16] This often replicated result has lead to the nearly universal interpretation that more frequent lexemes are somehow more psychologically active, and that sufficiently high frequency lexemes are able to be accessed through a more efficient process. Baayen (2010), on the other hand, has shown that the frequency of lexemes in the British National Corpus is very much predictable from a set of morphological and syntactic factors. Hence, the raw repetition of a lexeme is hardly a simple and irreducible phenomenon. Nevertheless, Baayen's study at the same time confirms that whole-word token frequency explains the largest proportion of the variance in lexical decision times. Consequently, word frequency may be thought of as a major principal component in the determination of lexical access speed. Whether this interpretation need necessarily modify the commonplace view that the rapid lexical access associated with high-token frequency lexemes is indicative of holistic morphological storage, requires further study.

lemma frequency as distinct from whole-word frequency, unfortunately, is relatively understudied; the vast majority of studies concerned with lexical token frequencies look to whole-word frequencies.[17]

Clear effects of simple absolute whole-word token frequency for lexical items have been reported across a wide variety of languages, using all manner of experimental modalities: higher token frequencies correlate with shorter response latencies and lower error rates in lexical decision tasks, for instance. Rather than report in detail here these findings, I refer the reader to the literature cited in Diependaele et al. 2012: 316–7. Worth noting is that Ford et al. (2003) found that, on visual lexical decision tasks with English speakers on nouns (for which the only inflectional options are singular or plural), whole-word frequency was the best predictor of response times; among English nouns, however, whole-word and lemma frequency correlate strongly, the ratio of lemma to whole-word forms not exceeding $3:1$. In other categories, such as English adjectives, where lemma frequency may markedly exceed certain whole-word forms (e.g., the comparative and superlative forms of adjectives), Ford et al. (2003) did find significant effects of lemma frequency.

Of greatest theoretical import, perhaps, is the fact that significant whole-word frequency effects are attested for inflected forms that belong to "regular" inflection patterns in a number of languages, including Serbo-Croatian (Katz et al. 1991), Finnish (Bertram et al. 2000), and Dutch (Baayen et al. 1997). This set of studies indicates that whole-word forms that ostensibly undergo identical inflectional processes may nevertheless have distinct psychological states. Moreover, such results would then flatly contradict the predictions of dual-mechanism models (cf. 2.3 above) of morphological processing that hold that forms instantiating "default" processes are always decomposed into a stem and inflectional morphemes; under such models, only lemma frequency effects would be predicted to exist for many fully inflected forms.

While fairly straightforward relationships appear to exist between whole-word frequency and lexical decision response latencies, the situation may be more complex with respect to lemma frequency. In particular, a study by Bien et al. (2011) on deverbal adjectival derivation and inflectional processes in Dutch reported a non-linear relationship between lemma frequency and response latency: the shortest response times were associated not with the highest-frequency lemmas, but lemmas of moderate frequency. But the experimental design of this study was such that moderate lemma frequencies were more common than either low or high frequency items, and in consequence, subjects were able to respond more quickly to lemma frequencies that were more likely, as subjects became attuned to the experiment-internal distribution of lemma frequencies.

### 3.2.2 Morpheme Token Frequency Effects: Root and Affix Frequency

Besides the frequencies of word forms or stems that may be associated with specific semantics, studies have also considered the role of token frequencies for individual morphemes,

[17] Distressingly, in some studies, whether whole-word or lemma frequency is the object under study is not always clear.

both roots and affixes. In languages such as English and Dutch, in which roots readily occur as simplex surface forms without any additional morphological components, the presentation of root items with higher token frequencies in corpora more reliably and substantially prime the recognition of their derivatives in lexical decision tasks. Furthermore, root token frequencies also correlate with smaller response latencies in lexical decision tasks generally; this result is particularly robust in non-prefixed forms presented auditorily. Again, these effects are sufficiently well replicated that the reader may refer back to the literature surveyed in Diependaele et al. 2012: 317.

One specific attempt to disentangle independent effects of frequency for roots, affixes, and whole words is a study of Italian by Burani and Thornton (2003). The authors ran three experiments, containing mixtures of both real Italian lexemes and pseudo-words generated from both genuine and pseudo-roots and genuine derivational suffixes; all experiments were lexical decision tasks with a visual mode of presentation. Experiment 1 tested reaction times using pseudo-words built with pseudo-roots and derivational suffixes of high (639–1557 tokens per 1.5 million words), medium (55–90 tokens), and low (7–18) frequency – higher suffix token frequency also appears to correspond to higher type frequency (mean type frequencies of 165, 20.6, and 6.3, respectively);[18] Experiment 2 employed genuine low-frequency words (1–10 tokens) exhibiting different combinations of roots and derivational suffixes of both high- and low-frequency; Experiment 3 likewise used genuine low-frequency (1–13 tokens) words of the four combinatorial types as in Experiment 2, but also included non-derived words with root frequencies matched to the frequencies of the derived words, containing roots of differing frequencies. The authors assume that, since all of the morphologically complex forms used in Experiments 2 and 3 have low whole-word frequencies, morphological segmentation would likely be necessary for processing.

In Experiment 1, high token frequency derivational suffixes appeared to cause significant interference with the successful identification of the form as a non-word: response latencies were significantly longer and error rates significantly higher with respect to the medium and low token frequency groups. This result suggests that higher token frequency suffixes are identified sufficiently quickly such that they ultimately interfere with the rejection of non-words. In real words containing high-frequency roots and high-frequency suffixes (Experiment 2), reaction times were both significantly faster and more accurate than in forms with either a low-frequency root or suffix, and were slowest and most erroneous, to a significant degree, for forms with both a low-frequency root and low-frequency suffixes. Experiment 3, however, which included items better controlled for familiarity (as rated by separate groups of subjects), found instead that forms with high-frequency roots patterned together in response times and error rates, while forms with low-frequency roots and non-derived words patterned together. The ultimate conclusion is that only root token frequency provides a significant processing advantage over non-derived forms; the token frequency of derivational suffixes seems less likely to play a substantial role.

[18] The procedure of grouping lexemes into frequency bins in such studies is commonplace, and simplifies certain kinds of statistical analysis of results (e.g., analysis of variance), though Ford et al. (2003) argue (I think rightly) that correlational and regression analysis using the actual frequencies and corresponding response latencies would be preferable.

### 3.2.3 Base-Derivative Relative Token Frequency

Hay (2001) (cf. also Hay 2003: Chs. 4, 5, & 6) has adduced substantial evidence to the effect that relative token frequencies between a morphologically complex form and its etymological base may be just as relevant to the processing speed of lexical access as absolute whole-word token frequencies. Generally speaking, in cases in which a morphologically complex form is of higher token frequency (e.g., *exactly* and *exact*, with respective token frequencies of $2535$ and $532$ in the CELEX corpus), Hay claims that the morphologically complex form will take on greater independence, because it maintains less association in the lexicon with its base. Hay's work, unfortunately, does not directly assess the relationship between relative frequency and word recognition, but instead uses a combination of meta-linguistic experimentation, semantic, and phonetic effects to argue that relative token frequency is of relevance.

As a meta-linguistic assessment, Hay presented lists of $17$ pairs of prefixed and $17$ pairs of suffixed English lexemes, in which one member of each pair had a token frequency greater than its base in the CELEX corpus (Hay 2001: 1046–50). Subjects made a forced-choice decision to identify which member of each pair seemed more "complex". Both by-subject and by-item analyses showed that forms with higher token frequencies than their bases were rated as the less complex item in the pair, with significant regularity ($p < 0.01$); overall, $65\%$ of datapoints analyzed found that the form with higher token frequency than its base was considered less complex. On the semantic side, Hay (2001: 1055–9) examined (in more detail in Hay 2003: 104–17) base-derivative relative frequency and the number of definitions listed for the derivative in Webster's 1913 Unabridged English Dictionary, thus considering the degree of polysemy of the derived forms. Hay ultimately concludes that polysemy itself is principally a factor of absolute token frequency – more frequent lexemes tend to develop more meanings – but derivatives with higher relative frequencies are more likely to lose their semantically transparent meanings.

Phonetic evidence for relative frequency effects appears in Hay 2003: 88–95 & Ch. 6. In one experiment, Hay evaluated the likelihood for a speaker to place a contrastive pitch accent on prefixed forms when reading sentences containing both a base and a semantically transparent derivative, as in the following example.

(8) a. Sarah thought the document was legible, but I found it completely illegible.
    b. Sarah thought her cousin was liberal, but I found him completely illiberal.

Contrastive sentences were interspersed with non-contrastive sentences in the experiment. In terms of results, Hay found that derivatives with token frequencies greater than their bases were significantly less likely to attract the contrastive pitch accent (by-item, $p < 0.02$), and furthermore, the greater the difference in (log-transformed) relative frequency, the greater the number of tokens subjects produced with a contrastive pitch accent.

A further study considered rates of /t/ deletion (e.g., in *softly*, *swiftly*, or *daftly*) as a function of base token frequency (Hay 2003: Ch. 6). Subjects were asked to read sentences containing forms with the suffix *-ly*; forms containing the sequence *-tly* were divided into three

groups: very low whole-word frequency, frequency of derivative less than frequency of base, and frequency of derivative greater than frequency of the base. The subjects were recorded, and the words of interest were spectrographically examined to measure the period between offset of a preceding segment and onset of /l/, to determine the duration of /t/. Low whole-word frequency forms (e.g., *daftly*) contained the greatest duration of /t/, while derivatives with frequencies less than their bases (e.g., *softly*) contained longer /t/ duration than in the case of the reverse relation (e.g., *swiftly*); indeed, for some forms in phonetic contexts particularly susceptible to /t/ reduction (such as *-ftly*), some subjects showed no greater /t/ duration in forms with frequencies greater than their bases than words containing no /t/ whatsoever (e.g., *briefly*).

Vannest et al. (2011) recently carried out a lexical decision task, also performing event-related fMRI measurements on subjects during the study, so as to track neurological effects of frequency during the task. The authors divided the English lexemes employed in the task into three groups: monomorphemic, "whole-word" (i.e., complex words containing less productive derivational suffixes, such as *-ity*), and "decomposable" (i.e., complex words containing more productive derivational suffixes, such as *-ness*); items in the "whole-word" and "decomposable" groups were matched for token frequency, base frequency, and family size (see 3.2.5), while monomorphemic items were of either high token and base frequency or low token and base frequency.[19] In terms of responses on the task itself, base frequency did not significantly affect either response latencies or error rates for words built with less productive processes (the "whole-word" group). In terms of fMRI measures, certain event-related signals in the left inferior and superior temporal gyri did not show an effect of base frequency, but of word type, with the "whole-word" group showing response times between the monomorphemic and the "decomposable" groups.

### 3.2.4 Type Frequency

Perhaps surprisingly, the type frequency of morphological processes is relatively under-investigated by psycholinguists with respect to effects on word processing rates.[20] In one such study, Laudanna et al. (1994) found that Italian prefixes with higher type frequency induced longer response latencies and higher error rates on pseudo-words in a visual lexical decision task, thereby indicating that affixes distributed over a greater number of different forms were more likely to be regarded as possible words. Instead, the role of type frequency has been assessed more often by linguists employing "wug" tests or rating tasks. Generally speaking, "wug"-test responses and high well-formedness correlate well with type frequency, thus indicating that, *ceteris paribus*, speakers are more likely to apply processes or accept novel items formed through processes that are instantiated in a large number of different

[19] However, for forms with low base frequency, the mean base frequency still exceeded mean token frequency. For instance, the mean token frequency of low base frequency "decomposable" used was $6.05$, mean base frequency for those forms was $28.08$. Thus, results here cannot be neatly considered in terms of relative frequency, though one might assume that low base frequency groups more likely contained forms for which surface token frequency may have exceeded base frequency.

[20] In their discussion of type frequency effects, Diependaele et al. 2012: 319 consider only experimental studies of family size effects, on which see 3.2.5 below.

words.

Albright 2002a and Albright and Hayes 2003 employed rating tasks with "wug" items with speakers of Italian and English, respectively, to evaluate the correlations between the performance of the Minimal Generalization Learner (cf. 2.3 above) and speakers' assessments of infinitives (Italian) and past tense (English) forms built to nonce stems. Since the MGL, as described above, calculates the confidence of a rule solely on the type frequency with which that rule applies or does not apply wherever its structural description is met in the training data, the extent to which its predictions align with rating task results should be indicative of the role of type frequency in processing and producing novel forms. In both the Italian study and Experiment 2 of the English study, subjects were auditorily presented with a sentence that would present a nonce verbal stem (in Italian, the 1.sg.pres., e.g., *lavesso*; in English, the stem in an infinitive, e.g., *to rife*), then with sentences containing different possibilities for the target class (in Italian, four different infinitives, one from each logically possible class; in English, a regular form with the suffix *-ed* and a form with root vowel change), and asked to rate each novel form on a scale from 1 to 7. After scaling the ratings data to the confidence scores generated by the MGL, significant correlations between the MGL predictions and participant ratings indeed systematically emerged (cf. Albright 2002a: 695–8 and Albright and Hayes 2003: 140–6, respectively).[21]

Dąbrowska (2008) carried out two "wug" production tasks with Polish speakers that elicited the inflection of nonce nouns for dative singular case. The task used a written format that introduced the noun in a nominative singular form in one sentence, then asked the subjects to fill in a blank in a following sentence with an appropriate form of that nonce word, by way of frame phrases that obligatory call for dative case in Polish. The nonce forms devised for the task were designed to belong to phonological neighborhoods of either high or low density. Based on Dąbrowska's source for frequency data, high neighborhood density nouns such masculine nouns ending in *-ator* or feminines ending in *-arka* had a mean type frequency of 232; all of the high density neighborhoods result from productive derivational suffixes. Low density neighborhoods, such as feminines in *-zia* or neuters in *-ro*, seem largely to share phonological shape in their final syllables by chance, and have a mean type frequency of only 21. Since the nouns belonging to neighborhoods largely or entirely select for the same dat.sg. ending, to identify them is essentially the same as speaking of "Islands of Reliability", in Albright's terms.[22] Across both experiments, subject performance was significantly

[21] Albright (2002a: 701) also calculated correlations with ratings data to an MGL learning run on the Italian data that calculated confidence based on token, rather than type frequency as per usual; Albright found that, in general, the type-based versions outperformed token based versions. Given the assumption that high token frequency forms may not really psychologically instantiate a morphological process at all (cf. 3.3 below), a version of the learner in which the contribution of a form either strengthening or weakening a rule's reliability diminishes with increasing token frequency of a given form, and, at the prediction phase, requires higher confidence rules to contravene a given high token frequency mapping, could be in order. In effect, for each input-output mapping, a form-specific rule could be learned that calculates confidence based on token frequency, and form specific rules could further compete with the more abstract rules generalized from multiple forms.

[22] In principle, the dat.sg. endings of Polish are largely determined by nominal gender: all masculines take the ending *-owi*, feminines take either (palatalizing) *-e* or *-i*, and neuters take either *-u* or are, very rarely, in-

more accurate (i.e., conformed to the target ending type for a given stem-final sequence) for high-density neighborhoods than low density neighborhoods; performance for many subjects was at ceiling for masculines and feminines.[23] This result supports the conclusion that the greater type frequency of specific morphological patterns supports the successful access and application of that process under novel circumstances.

### 3.2.5 Family Size Effects

The effects of a lexical item's family size on morphological processing seem first to have been investigated in Schreuder and Baayen 1997, where family size was measured as the sum of all distinct derivatives and compounds in which a monomorphemic noun occurs. Experiment 3 in that study found that, for monomorphemic Dutch nouns with roughly equal cumulative family frequencies (i.e., the token frequency of a root in all derivatives and compounds), lexical decision responses were significantly faster and more accurate for nouns with larger family sizes. Analyses in Ford et al. 2003: 101 on family size of English monomorphemic nouns also found that family size counted as all derivational forms and all usage in compounds (so for instance, a primary derivative and a root compound would each count as a family member) accounted for more variance than any other way of calculating family size, and more variance than a single variable made from collapsing all family size effect variables into a single variable through principal components analysis; this result supports the metric of family size first employed by Schreuder and Baayen (1997).

de Jong et al. (2003: 66) argue that the family size effect is, at root, a semantic effect, for three reasons: 1) the effect appears in lexical decision tasks (where participants see or hear the entire word immediately), but not in progressive demasking tasks (where the correct reading and interpretation of the word based on available exposure is necessarily gradual); 2) semantically distant or opaque family members do not contribute to the effect – correlations between response latencies and family size are better when opaque items are removed from the data set; 3) de Jong et al. (2000) have shown that family members with morphologically induced phonological differences (such as the past tense stem or past participle of strong verbs in Germanic languages) also contribute to the effect, thereby showing that the family size effect is not (solely) an effect of phonological or orthographic similarity. Words which are phonologically or orthographically closer, but semantically unrelated (e.g., Dutch *vocht* 'moisture' is not connected with the past participle *gevochten* of *vechten* 'fight'), do not experience faster response latencies from those semantically unrelated forms. Furthermore,

declinable (having a single form for all case functions). Hence, insofar as the form of the noun could reliably indicate gender, the inflectional ending is largely deducible.

[23] In fact, subject-level differences in performance were highly correlated with number of years of education, and more educated speakers performed at ceiling for masculines regardless of neighborhood density, and at ceiling for one of two classes of feminines regardless of neighborhood density. Hence, the effect of neighborhood density was clearer among subjects with fewer years of education, for whom neighborhood density clearly affected performance. Dąbrowska attributes the marked difference in performance between more and less educated subjects to two factors: larger vocabulary size (assessed through another test of the subjects) and greater experience with "archaic" or "high-register" dative constructions in written texts that provide more knowledge of neuter nouns in the dative.

del Prado Martín et al. (2005) found that family size might instead be inhibitory for Hebrew speakers when a given root participates in several different semantic fields; this finding suggests that family size should perhaps be separated into a set of items with a transparent semantic relationship to a given form, and a set with an opaque semantic (but still clear formal) relationship to a given form.

The study of del Prado Martín et al. (2005) is also very much applicable to the relevance of family size for old Indo-European languages, in which roots often occur with many different derivational suffixes, but never appear as simplex nouns or verbs to which inflection directly applies (see further 4.1). The authors there establish that family size for Hebrew, a language replete with non-concatenative morphological derivations, is best measured as the total number of distinct lexemes in which a given consonantal root occurs. However, del Prado Martín et al. 2004 determined that, for Finnish speakers, only the number of lexemes *derived from* a given lexeme, but not the base of or other relatives of a lexeme, appear to count towards the family size. For instance, for the lexeme *työläinen* 'worker', a derivative *käsityöläinen* 'craftsman' and the further compound *käsityöläinenmuseo* 'handicraft museum' count towards the family size, but not the base *työ* 'work' or related derivatives of that base, such as *työkalu* 'tool' or *työläs* 'laborious'. A reanalysis of some earlier experiments concerning family size in Dutch also revealed that family size was better measured for Dutch as in Finnish. On balance, family size in a language such as Sanskrit should probably be calculated as for Finnish, but the matter remains uncertain.

While the papers discussed thus far deal exclusively with derivational family size, Traficante and Burani (2003) explicitly studied *inflectional* family size (i.e., the richness and extensiveness of an inflectional paradigm) on Italian data. Although the Italian verbal system does not exhibit as great a degree of complexity and as many possible forms as Sanskrit and Ancient Greek, it is considerably more complex than English, or even present-day German or Dutch. Traficante and Burani (2003) conclude that a large inflectional family favors parsed access, whereas a small inflectional family tendentially applies whole-word access to its members. Given that both the nominal paradigms and verbal paradigms of Greek and Sanskrit, are considerably more complex than Italian adjectives, their members may typically be subject to parsed access in inflection as well.

### 3.2.6 Phonotactic Probabilities

In addition to the frequencies of words and morphemes themselves, work by Hay (2003: Chs. 2 & 3) has argued that phonotactic probabilities may impact morphological processing as well. In a forced-choice experiment, English speakers presented with pairs of disyllabic nonsense words (e.g., *vilfim* and *vipfim*) were significantly more likely to rate as complex the member of the pair which has a lower probability coda-onset transition (in the CELEX corpus, [lf] occurs 11 times, while [pf] never occurs). Hay further identified, within the English lexicon, a number of interesting correlates of affixed lexemes that contain low-probability junctural phonotactics at the root-affix boundary: bases and derived forms are rated as "less related" (Wurm 1997), the derived form exhibits greater polysemy, and the derived form has a greater likelihood of having higher token frequency than its base. Hay (2007) presents fur-

ther phonetic evidence: based on a spoken corpus of New Zealand English, Hay found that, in words beginning with the sequence *un-*, greater length of that first syllable correlates with rarer consonantal phonotactic transitions (e.g., it is longer in *uncork* than *unhinge*, where the sequence [nh] is more common), in addition to evidence for substantially greater phonetic reduction of the *un* in morphologically simplex lexemes (e.g., *unless* or *until*).

### 3.2.7 Summary

The array of possible determinants of morphological processing seems nearly overwhelming. On the whole, almost certainly all of the effects described above play a role of greater or lesser significance in the process of lexical access for morphologically complex words. Although they are referring specifically to response times in lexical decision tasks, the remarks of Ford et al. (2003: 114) could apply to the question of morphological access more generally: it is "not simply the result of the resting activation level of a single representation, but rather a result of a decision process, in which a number of sources of information, form [i.e., phonetics and phonology], morphology, and semantics are utilized. These sources of information can clearly be subdivided further, with morphological information comprising the various frequency/family size variables that have been shown to affect response times."

Nevertheless, some clear general trends concerning the effects of frequency appear sufficiently robust that their impact may be stated plainly here:

- greater whole-word token frequency increases processing speed (3.2.1).
- greater lemma token frequency increases processing speed, though for lemmata with few inflectional variants, the effects of whole-word token frequency may readily mask lemma frequency (3.2.1).
- at least for low token frequency whole words, higher root token frequency increases processing speed (3.2.2).
- morphologically complex forms with higher whole-word token frequencies than their bases tend to become less semantically transparent and more phonetically reduced, to an extent that whole-word token frequency alone cannot explain (3.2.3).
- processes or affixes with higher type frequency (modulo "Islands of Reliability" or "neighborhood density") produce novel forms that speakers find more acceptable (3.2.4).
- the number of derivatives related to or derived from a given form tends to speed processing as a semantic effect, though some differences between languages may exist (3.2.5).
- low probability phonotactic transitions correlate with perceived greater complexity and less phonetic reduction (3.2.6).

Despite the wealth of data and evidence available, Plag (2007: 199) is nonetheless correct to emphasize the point that such experimental results are always subject to interpretation

and possible reanalysis, and that, moreover, we lack a model that comprehensively integrates all of the results of and claims made by all of the studies that I have surveyed here. However, I believe that the greater part of these effects find a reasonable interpretation under a *dual-route* model of morphological processing, which allows for both full-form (holistic) access and parsed access for items at all levels of morphological complexity: an entry in the mental lexicon is possible at the level of the whole word, including inflection, or a stem, including any derivational processes, or as a set of distinct morphological elements. Just such a model will also allow for a plausible interpretation of morphological productivity.

## 3.3 The Morphological Race Model and Productivity

An adequate psycholinguistic model of morphological processing and production should address both the experimental facts that bear on the speed with which forms are processed, as well as be able to account for why productive derivational and inflectional processes are preferentially selected and applied in production (and on the processing side, why non-productive applications, such as an Eng. **stepth* are recognized as impossible or ridiculous). Some models represent the extremes of logical possibility: Butterworth (1983) proposes that all existing known full word forms are always retrieved from memory in that full form and that morphological processes then apply only for the production of novel or unknown forms, and never for processing; Taft and Forster (1975), conversely, argue that all morphologically complex forms (however they might be recognized as such), are always necessarily decomposed into their constituent elements, at all times. A storage-only model easily accounts for the robust effect of whole-word token frequency in lexical decision tasks, but in fact does not and cannot really say anything meaningful about the effects of type frequency, or explain why an apparent process would be preferentially selected in production (i.e., productivity).[24] Conversely, parsing-oriented models would predict more robust effects of root and morpheme token frequency, since those units would be the levels of access, instead of a dominant role for whole-word token frequency.

Both Baayen (1992: 125–33) and Frauenfelder and Schreuder (1992) recognize that a systematic accounting of the facts pertaining to token frequency, type frequency, and productivity requires a substantial role for both the storage and the parsing component in the process of lexical access. Frauenfelder and Schreuder sketch a Morphological Race Model (MRM), the basic mechanics of which are as follows: in lexical processing, a morphologically complex word may be accessed either as a whole word, retrieved from a single entry in the lexicon (a *holistic* route), or from a combination of roots, stems, morphemes and word-formation and inflectional processes that exist in the lexicon (a *parsed* route). These two possible routes of lexical retrieval operate in parallel, but the route by which the word is most quickly retrieved is said to "win" the race. However, both routes contribute to, inter-

[24] Note that the production aspect of a storage-only model would look rather like exemplar-based models, such as MBL, discussed at 2.3 above. The issue thus really returns to the plausibility of a massive lexicon-wide computation that synthesizes all of the generalities therein, just in case a new form needs to be produced, as opposed to a gradually reinforced abstraction (à la "rules" in MGL).

act, and may speed the overall access of a given lexeme. Consequently, this model predicts that, among two lexemes of equal whole-word token frequency, the form belonging to a more productive category will be accessed more quickly, because the parsing of a more productive process will be able to speed lexical access more than a less productive process.

How quickly any elements may be accessed or parsed depends upon their "resting activation level" for a morphological process ( Baayen (1993) calls this $\mathscr{A}$). Frauenfelder and Schreuder (1992: 176) say that "the resting activation levels of access representations of the stem and affix will be increased *only* when the parsing route wins the race and produces a *successful* parse. A successful parse is one in which the analysis of the stem and its affix(es) [or, presumably, morphological processes more generally – RS] leads to a meaningful interpretation." Baayen and Schreuder (1995: 133–6) more explicitly define the process of parsing as involving three stages (cf. Figure 3.1): phonology and segmentation, licensing, and composition. These three stages approximate the interfaces between morphology and phonetics/phonology, syntax, and semantics, respectively. Phonology and segmentation create access representations, which are licensed by the syntax and assigned meaning by the semantics through an intervening concept node.[25]

For the question of modality of lexical access, what is most relevant is the availability and strength of a full-form access representation relative to the strength of the access representations for individual segmentable components, as well as the availability and need for an independent concept node linked to a given access node. The overall resting activation level for given morphemes flows not only forward from access representations to concept nodes and syntactic/semantic nodes, but back from the syntax and semantics to the concept and appropriate access nodes. Consider the contrast between Eng. *disease* and *discomfort*: since the latter is semantically compositional, some of the activation of the semantic nodes activated will flow back ultimately to the access representations /dɪs-/ and /kʌmfərt/, even if the full-form access representation principally activated the appropriate concept and semantic nodes; conversely, since the former is non-compositional, very little, if any activation will flow back to the access representations /dɪs-/ and /iːz/, but almost solely to the representation /dɪziːz/.

We are now in a position to make sense of the psycholinguistic results surveyed in the preceding section, which in turn allows for a sensible interpretation of why $\mathscr{P}$ (cf. section 2.2.2) provides a good approximation of productivity. First, the faster processing speed that obtains at the whole-word, lemma, and root levels from greater token frequency follows from greater resting activation levels at both access representations and concept nodes that accumulate with repeated exposure. However, while lemmas and roots will have a greater token frequency than whole words, the fact that lemmas and roots require segmentation of inflectional and derivational morphology to occur in order to activate their access representations renders the relevance of their token frequencies dependent upon how readily segmentable the overall morphology of the whole word is. Root and lemma frequency effects also reflect increased activation of the same concept node that will be shared by a number of access rep-

[25] The segmentation and licensing aspects of the model are computationally implemented in in Baayen and Schreuder 1999; the whole model is also basically accepted in Zuraw 2009.

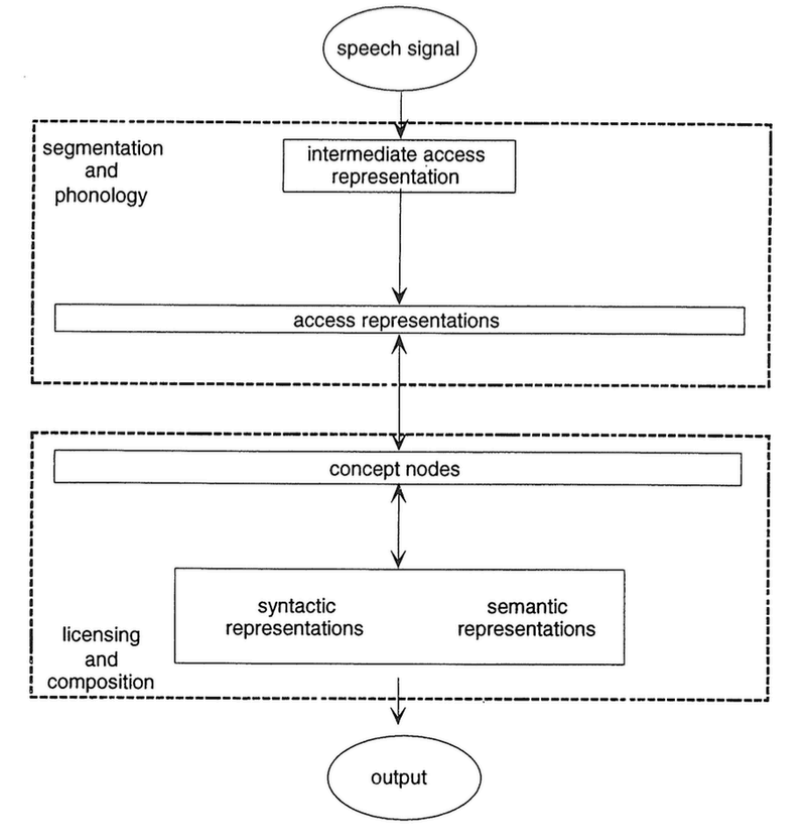



Figure 3.1: General Structure of a Model of Morphological Processing (from Baayen and Schreuder 1995: 134)

resentations (e.g., the access representations /beɪkərz/, /beɪkər+s/, and /beɪk+ər+s/ all benefit from activation of the concept node BAKE). Thus, lemma token frequency will play an important role for infrequent paradigm members, but not for very frequent paradigm members, for which whole-word access would be more efficient. Similarly, the greater relevance of root token frequency in pseudo-words follows from the interpretation that a pseudo-word, like a novel form, would have neither a whole-word nor a lemma access representation, and require access via full decomposition, where a more frequent root's access representation and concept node will become available more quickly.

As an example, consider the token frequencies of the Skt. lemma *ūtí-* 'aid' (root $\sqrt{av^{i}}$ 'help' + derivational suffix /-ti-/), which occurs $350\times$ in the RV, with wildly different frequencies for its paradigm members: dat.sg. *ūtáye* and inst.pl. *ūtíbhiḥ* each occur $100\times$, while nom.sg. *ūtíḥ*, acc.sg. *ūtím*, and loc.pl. *ūtíṣu* occur only $6\times$, $3\times$, and $1\times$, respectively; whole-word access representations /uːtájeː/ and /uːtíb̥is/ likely dominate for the dat.sg. and inst.pl., while parsing of the inflectional morphology and access via the representations for the lemma /uːtí-/ and the acc.sg. morpheme /-m/ (or even root /əw$^{i}$-/ + suffix /-ti-/ + inflection /-m/) seems more likely for *ūtím*.

The semantic effect represented through family size, meanwhile, follows from the feedback to the concept nodes and access representations in activation of semantic nodes (cf. Schreuder and Baayen 1997: 132–4): the greater the number of semantically related word-forms that exist in the lexicon, the greater the regularity with which some activation of specific concept and access nodes will flow back from the activation of the semantic node that closely binds together the members of the family. In the same fashion, Hay's finding that derivatives having greater token frequency than their bases become more polysemous and less semantically transparent follows from the greater number of opportunities that the high-frequency derivative has to link with other semantic nodes, and in virtue of those connections, will less often be semantically re-centered through semantic feedback that comes when its base is used. Diachronically speaking then, the semantic drift of a high-frequency derivative away from its base will also necessitate access through a holistic access representation, since parsed access will not associate to the appropriate concept node that mediates the correct semantic interpretation.

The role for hapax legomena in the interpretation and measurement of productivity, under this model, is entirely clear: because the genuine neologisms that hapaxes approximate would necessarily undergo parsed processing or production, their rate of occurrence sets some minimum number of tokens that increase the activation of morphemes and morphological processes. The precise place of type frequency itself, on the other hand, is somewhat ambiguous within the MRM. If Frauenfelder and Schreuder (1992), as cited above, are correct that only the cases in which parsed access "wins" a race (i.e., the component access representations all come on-line more quickly than the whole-word access representation), then what high type frequency really reflects is a comparatively high number of tokens of low token frequency forms. That is to say, high type frequency morphemes and processes may be preferentially applied and rated highly by speakers because they happen to contain substantial numbers of forms that require parsed access.

Generally, the processes with highest confidence, in MGL terms, because they contain larger numbers of types, then also contain a greater number of low token frequency items for which the application of segmentation, on the processing side, would be necessary; on the production side, it reflects the greater resting activation of a process conditioned on some input. Perhaps what the rules induced by the MGL really reflect is the productivity of the morphological process conditioned over a phonological context, i.e., the extent of the activation of a context-conditioned process. From the processing side, it reflects what processes need to be applied at the segmentation stage: just as the "undoing" of the (productive) phonological process that generated an anaptyctic vowel in Eng. [baɪdəd] is necessary to obtain the access representation /baid+d/, so too would the morphophonological process represented in the Gk. sigmatic aorist δουλωs- [do:lɔ:s-] 'enslave, become a slave' (reflected in the construction in the example below; cf. examples (4) and (6) in Chapter 1) need to be "undone" in order to recover the morphemes /do:lo-/ and /-s-/.

(9) $< [\ [\ldots \text{V:}]_{\text{Vj}}\ [\text{s}]_{\text{k}}\ ]_{\text{i}} \leftrightarrow [\text{action of } \text{V}_{\text{j}} + \textsc{past.perfective}_{\text{k}}]_{\text{i}} >$

We might hypothesize that, humans always possessing the capacity for abstraction over their input data, will be most susceptible to the effects of type frequency when their data is

relatively impoverished (i.e., for young children and L2 speakers): comparatively few types have sufficiently many tokens to become entrenched, and thus susceptible to holistic access, while substantial numbers of types will be of sufficiently low token frequency that the parsed access option and its morphemes are regularly activated. Presumably, with exposure to more tokens the need for parsed access is lessened, and entrenchment of whole word forms sets in (hence the pattern of overgeneralization errors in morphology in L1 acquisition that gradually die out, as amply documented in, e.g., Marcus et al. 1992).

Thus the precise role for type frequency in morphological processing, to my knowledge, remains an empirical question. Perhaps Frauenfelder and Schreuder (1992)'s formulation, that only "winning" parsed forms are of relevance is too strong, and rather, that even forms which are ultimately accessed holistically contribute partially to the activation of parsed processes, in proportion to the speed of the respective modes of access. On this point, diachronic data could be of service, in comparing the performance of processes with inverse ratios of $\mathscr{P}$ and $V$.

## 3.4 Correlates of Productivity Measures: Hay and Baayen 2002 and 2003

Hay and Baayen (2002) and Hay and Baayen (2003) proceed from the foundation that, as discussed above, the number of "parsed" or "decomposed" items in the lexicon for a process (i.e., those words that have a lexical representation that calls upon the component morphemes or morphological processes for the meaning of those words) is directly linked to morphological productivity. In effect, words that are processed in perception as morphological simplexes do not contribute to the psychological activation of a morphological process, whereas words that are processed as being morphologically complex will activate the necessary morphological processes, since those processes were used to understand the words.

Hay and Baayen (2002) attempt to determine more exactly the frequency ratios that would entail parsed lexical access, as well as establishing whether in fact productivity in production correlates systematically with parsing in perception. Hay and Baayen refer to the relative token threshold below which a word would be parsed as the "parsing line"; the process by which they determine this "parsing line" is through the MATCHECK model described in Baayen and Schreuder 2000.[26]

The important conclusion that emerges from the establishment of a "parsing line" is that, seeing that this line is greater than $x = y$ (where $x$ is derived frequency and $y$ is base frequency), whole-word access in general has an advantage over parsed access, even for words that have a frequency slightly less than their bases.[27] Consequently, the derived words that have token frequencies that are notably less than their bases, among which hapax legomena and true neologisms would likely fall, are the best candidates for parsing in perception.

[26] A implementation of MATCHECK in Perl by Kie Zuraw is available at `http://www.linguistics.ucla.edu/people/zuraw/prosword_2006/matcheckI.txt`.

[27] In principle, different parsing lines exist for each different morphological process, but, as a heuristic, Hay and Baayen (2002: 15) adopt the parsing line for the English suffix *-ness* (slope of .76, intercept of 3.76) as a general parsing line.

Having determined a parsing line, Hay and Baayen are then able to establish "parsing ratios" for the (English) morphological processes that they investigate: the proportion of words belonging to a morphological category that fall above the parsing line, and for which parsed access plays a considerable role. Crucially, Hay and Baayen find that the parsing ratio for a process has a strong statistical correlation with the $\mathscr{P}$ value of that process. This correlation between $\mathscr{P}$ and parsing ratios supports the hypothesis that, because parsing contributes to the psychological activation of a morphological process, processes that contain a considerable number of members that undergo parsed access will tend to be more productive. In effect, parsing ratios reflect in perception what $\mathscr{P}$ reflects in production.

Meanwhile, the total number of forms that fall above the parsing line for a process (rather than the *proportion* that fall above the parsing line) exhibit significant correlations with $\mathscr{P}*$ and type frequency ($V = \mathscr{U}$), rather than with $\mathscr{P}$. The correlation between the number of hapaxes that instantiate a process and the number of parsed tokens

> ...is by no means trivial. Hapaxes contribute extremely minimally to overall token counts, and so there is no a priori reason we should expect the number of hapaxes to correlate with the total number tokens which are parsed. Yet we *do* see this relationship, and the reason we see it (we suggest) is because there is a causal relationship between parsing and productivity. The larger the number of tokens that is parsed, the more activated and robust the representation of the affix is, and so the more available it becomes for the formation of new forms.
>
> Hay and Baayen 2002: 26

In sum, this research on lexical parsing serves to validate $\mathscr{P}$ and $\mathscr{P}*$ as measures of productivity: $\mathscr{P}$ is a function of the overall likelihood that a given process is subject to parsing, while $\mathscr{P}*$ is a function of the frequency with which a given process is activated (cf. Hay and Baayen 2002: 30).

Hay and Baayen (2003) go yet further, and seek out factors that could potentially impact parsing in perception besides relative token frequency. In principle, the authors suggest that "any factor which is involved in the segmentation of words from running speech also plays some role in affecting morphological decomposition" (pg. 7). Since such factors are multifarious, Hay and Baayen restrict their study to phonotactics (building on the work of Hay 2003 surveyed above). The potential importance of phonotactics in morphological analysis comes to the fore in the study of Raffelsiefen (1999), who builds upon the observation that English "level two" suffixes (in the framework of Lexical Phonology and Morphology, after Kiparsky 1982a) are predominantly consonant-initial, and so claims that the genuine division in English derivational suffixes lies in the division between consonant-initial and vowel-initial suffixes. This effect may result from the fact that infrequent-word internal junctures (which more often involve consonant-initial suffixes) are more similar to the junctures that occur between words; in effect, a phonotactically unusual juncture between a stem and affix could given the impression of a word boundary, which would act as a signal to parse the stem and affix just as two distinct words would be parsed.

In the course of this study, Hay and Baayen find that $\mathscr{P}$ and the token frequency of a morphological process are significantly correlated with the mean probability of the (rare) phonotactic junctures in the language that the process creates; $\mathscr{P}*$, meanwhile, shows a significant correlation with the number of types in the process that contain an illegal juncture (i.e., junctures that never occur in monomorphemic words of the language), and type frequency exhibits significant correlations with both the mean probability of juncture and number of types with an illegal juncture (though not a strong as $\mathscr{P}$ and $\mathscr{P}*$ respectively show). Just as in the 2002 paper, $\mathscr{P}$ is correlated with a perceptual factor that seems to concern just the morphological category itself, whereas $\mathscr{P}*$ is correlated with a factor that has global relevance in the sample. Rare phonotactic sequences facilitate parsing and thus productivity.

The clustering effect observed between the relative frequency ratios, junctural phonotactics, and the measure $\mathscr{P}$ gains an potential explanation in the observation that uncommon junctural phonotactics correlate with frequent base words (Hay and Baayen 2003: 27). Precisely because a multimorphemic word shows rare phonotactics, its probability of being parsed increases. Conversely, complex words having common phonotactics are more likely to be accessed as whole words (monomorphemically); this whole-word access may in turn facilitate semantic separation and drift from the etymological base, which in makes the words semantically opaque, and hence even less accessible via parsing. In effect, rare junctural phonotactics help to preserve parsing effects, and likewise to maintain semantic transparency, and bolster base frequencies that exceed derived frequencies. $\mathscr{P}*$ and type frequency apparently cluster together due to brute frequency of usage: a high number of types that instantiate a process will tend to maintain the psychological activation of that process. This activation is all the more effective if there is a considerable number of unique words instantiating that process, and if a considerable number of words contain illegal junctures.

Hay and Baayen (2003: 28 ff.) classify the basic division between what $\mathscr{P}$ and $\mathscr{P}*$ measure in terms of productivity as a difference between "parsability" and "usefulness": $\mathscr{P}$ reflects processes that are highly "parsable" while $\mathscr{P}*$ is indicative of highly "useful" processes. Morphological processes with high parsability maintain productivity because they are consistently instantiated by numerous parsed forms; processes with high usefulness may have fewer parsed forms overall, but their persistent frequency in the lexicon may cause some degree of activation, along with the number of forms in that group that are indeed parsable.

The two studies that I have summarized in this section I believe basically validate $\mathscr{P}$ and $\mathscr{P}*$ as measures of morphological productivity. The statistical correlation of these measures with relative frequency effects (in Hay and Baayen 2002) and phonotactics at morpheme boundaries (in Hay and Baayen 2003) demonstrates that these two productivity measures are grounded in effects emergent from the lexicon. That is, the distribution of words, their relation to other words, and some properties of those words, all plausibly affect how a language user interprets and further employs words. Hence, the distribution of forms in a language corpus, if conceived of as primary linguistic data, should be predictive of the kind of generalizations that a learner would make from that data. $\mathscr{P}$ and $\mathscr{P}*$ specifically reflect the probability that a speaker, having been trained on that particular corpus data, will use

some morphological process, in accordance with the definition of productivity developed in Chapter 1.

## 3.5 Conclusions

In this chapter, I have mustered ample evidence behind the claim that the tracking of frequency information is a fundamental and low-cost cognitive operation, the availability of which is necessary to account for key aspects of animal behavior and human psychology. Given the basicness of frequency information to cognition, the relevance of frequency to language is wholly unsurprising. Of greatest relevance for the present enterprise are the multitude of interacting frequency effects that co-determine the processing of morphologically complex words, and in particular, how those frequencies impact the resting activation levels of access representations, thereby leading to the differences in performance for different lexemes that are attested experimentally. Finally, the good correlation demonstrated to hold between the metrics of $\mathscr{P}$ and $\mathscr{P}*$ with many of the basic word frequencies and relative word frequencies strongly suggests that those metrics indeed capture some of the psychological realities that decide the productivity of morphological processes.

Nevertheless, while a metric like $\mathscr{P}$ may provide an effective summarization of how parsable a category is on average, or how likely that category is to take on new members, it does not entirely determine the parsability of an individual type. Since both grammatical behavior at the synchronic level and change diachronically manifest themselves in individual lexical items, discussion and analysis of frequency effects on morphology will still require the bouquet of frequency data surveyed in chapter. In short, because all of the frequencies discussed potentially have some relevance, none can be dismissed *a priori*, and ideally, should all be taken into consideration in a model of a process' productivity.

# CHAPTER 4

# Practical Preliminaries: The Languages and Corpora under Study

The function of this chapter is to provide some background information on the languages and corpora that will be the focus of the case studies in Part II. The two principal languages here are Ancient (specifically Homeric) Greek and Vedic Sanskrit. Since my interest is to discuss the productivity of derivational formations in the noun and verb, I first describe the major features of derivational morphology that Greek and Vedic share. I then proceed to describe the nature of the particular corpora that I will exploit in my studies.

Given that the studies in Part II are based on two natural-language corpora, rather than plain grammatical patterns, forms, or utterances stemming from my own knowledge, grammars and dictionaries, or speakers of the languages, it is basically a work of corpus linguistics. Consequently, some techniques that have developed in order to treat corpus data in linguistic research will necessarily be brought to bear. At 4.3, I therefore discuss whether naturally occurring frequency data of corpora can and should be transformed in order to obtain more reliable results, as well as the problem of how to regard and classify morphologically complex words that may instantiate more than one derivational process.

## 4.1 The Languages: Basics of Morphology in Old Indo-European Languages

Since this dissertation is concerned principally with aspects of morphology in two old Indo-European languages, a sketch of some common features shared in the morphological profile of old Indo-European languages is in order here. For a reliable overview and further details of the essentials of the morphological system in Proto-Indo-European and its daughter languages, see the relevant chapters in Fortson 2010.

The characteristics of inflectional and derivational morphology are, in large measure, similar across the oldest attested Indo-European daughter languages. The description here rests mainly on morphological features more or less shared between Anatolian (mostly as represented by Hittite), Old Indic (Vedic Sanskrit), Old Iranian (Avestan and Old Persian), Ancient Greek, and Italic (mostly as represented by Latin). In terms of lexical classes, these languages possess nouns, pronouns, and numerals, adjectives, verbs, pre- and postpositions, and assorted discourse particles; a well-defined class of adverbs does not exist in most of these languages, adverbial functions usually being fulfilled by case forms of nouns and adjectives, or prepositions. Pre- and postpositions and particles normally have unchanging form (i.e., they are not declined or inflected), though they may participate in some word-

formation processes. In particular, all of these languages exhibit a greater or lesser tendency for prepositions serving as preverbs to become incorporated into the verbal lexeme proper, thus creating prefixes, which in turn can be used derivationally.

Every noun and verb in these languages can principally be analyzed into three components: a root, a derivational suffix, and an inflectional ending. The template in example (10) reflects the usual extent of a nominal or verbal formation.

(10) ⟨Prefix⟩ + Root + (**Suffix$_1$**) + (Suffix$_2$) + Inflectional Ending

The angled brackets around the prefix slot reflect the relative rarity and less obligatory nature of prefixing morphology (excepting reduplicants) in the oldest IE languages; whether any fully grammaticalized prefixes existed in PIE is debatable.[1] While not, strictly speaking, obligatory, I bold the **Suffix$_1$** slot because the vast majority of both nominal and verbal forms contain minimally one derivational suffix, at least etymologically. In many of the languages, the boundary between derivational and inflectional suffixes (e.g., in the high type frequency Greek and Latin second declensions) is somewhat opaque, though in such cases one could speak of the productivity of a derivable inflectional pattern, if not necessarily a distinct derivational suffix.

The root is the semantic core of the word; derivational suffixes determine inflectional classes for nouns and tense/aspect/mood for verbs; endings mostly convey morphosyntactic information. Those nominal and verbal forms that consist solely of a root and endings, without any derivational affix, are termed ROOT FORMATIONS. In the grammatical description of Old Indic and Old Iranian, as well as old Indo-European linguistics, the root itself is usually regarded as the base of derivation. Thus, for example, the Skt. root $\sqrt{piṣ}$ 'crush' is said to build a nasal-infix present, *pináṣṭi* (3.sg.pres.act.ind.), a sigmatic aorist, *ápikṣan* (3.pl.aor.act.ind.), and a perfect *pipéṣa* (3.sg.perf.act.ind.); similarly, Hitt. *pata-* 'foot', looks to contain a root *pat-*, and inflects according to nouns derived with a suffix *-a*. All of these formations are PRIMARY derivatives, because they are not built to any other existing surface stem found in Old Indic and Hittite, respectively; if a root aorist $^{X}$(*á*)*peṭ* 'crushed' (3.sg.aor.act.ind.) did exist, then the nasal-infix present and the perfect could be considered SECONDARY derivatives derived from the root aorist stem;[2] similarly, many nouns with the suffix *-a* are transparently derived from verbs that synchronically exist in Hitt., e.g., *karša-* 'a cut' from *karš-* 'cut'. The distinction between primary and secondary derivatives is crucial to an understanding of morphology in Greek and Sanskrit. Nearly all forms with suffixes belonging to the Suffix$_2$ slot are secondary derivatives, possessing a synchronically available base; many forms showing only a suffix in the Suffix$_1$ slot are primary derivatives, and appear to lack a clear synchronically available base.

[1] Though already in Homer, one can see that about 18 derivational prefixes originating as prepositions have developed, and, would seem to be very productive. The same tendency for the development of prefixes out of adpositions is observable in many Indo-European languages.

[2] Indeed, Indo-Europeanists regularly assume that, historically, all formations containing a derivational suffix rest upon some original root formation; such hypothetical root formations presumably were replaced by other derived formations, or fell out of use altogether.

### 4.1.1 Nouns

In all of the languages listed above, nouns and adjectives are inflected for case and number. All languages distinguish singular and plural in number, and all but Anatolian show a dual number at least residually, although use of the dual, as attestation shows, was most common with natural pairs (e.g., Skt. *ákṣi* 'two eyes').

Nouns inherently belong to to noun classes (genders), and the possibility of merely changing gender through an derivation exists (e.g., Av. *nar-* 'man' : *nairī-* 'woman'). Adjectives follow the gender of the nouns that they modify, which may entail a corresponding change in derivational suffix or inflectional pattern. By and large, nominal derivational suffixes build nouns of only certain genders (e.g., *a*-stems in Skt. build only masculines and neuters, while *ā*-stems build only feminines). The number of different nominal derivational suffixes is considerable, and the precise function of many suffixes remains uncertain. This uncertainty is more endemic to derivational suffixes that occur mainly as primary derivatives (i.e., in the $\text{Suffix}_1$ slot), than to suffixes that are largely susceptible to analysis as secondary derivatives (i.e., in the $\text{Suffix}_2$ slot).

Assorted nominal derivational affixes and their functions are well-documented in grammars of the relevant languages (e.g., Wackernagel and Debrunner 1954 for Vedic, Schwyzer 1938 [1953]: 455–543 for Greek, Hoffner and Melchert 2008: 51–63 for Hittite). These grammatical sketches of word-formation in the noun give a representative collection of types for the various affixes, but do not claim comprehensiveness in type counts, and certainly give no word frequency information. Specialized monographs on particular derivational affixes (e.g., Benveniste 1935 on infinitives in Avestan, Meissner 2006 on *s*-stems in Greek), where available, offer more complete collections of data.

### 4.1.2 Verbs

More so than in the discussion of word-formation in the noun, treatment of verbal stem formation in old Indo-European languages centers around abstract roots, with various ablaut patterns and derivational suffixes. Just as with nouns, the verbal stem may be coextensive with the root (i.e., no derivational suffix is found); such stems are root formations, and hence the terms 'root present' and 'root aorist'. In such cases, the root is less abstract, because the possible lexical entry for the root is mediated in surface forms only by inflectional affixes. More common, however, is to find the tense/aspect/mood stem of a verbal stem built through an overt derivational process. This process is usually a derivational suffix, though reduplicative processes exist as well, and are especially numerous in Sanskrit.

The verbal systems of Indo-Iranian and Greek are fundamentally the same in terms of overall architecture and their specific morphological components. A verbal stem basically expresses aspect: imperfective (traditionally termed "present"), perfective (traditionally termed "aorist"), or attained state, further developing to anterior (traditionally termed "perfect"). In the mapping of aspect on to tense, stems with imperfective aspect ("present") may map to ei-

ther past or present tense, while perfective stems ("aorist") map only to past tense.[3] Aspect is a derivational category in these languages. Inflectional endings mark person, number, and voice (active or middle). Indo-Iranian and Greek share a cognate valency-increasing derivational affix (causative, though this is largely lexicalized in Greek), a cognate past tense inflectional prefix (traditionally termed the 'augment'), and both have independently grammaticalized valency-decreasing derivational affixes (passive). The role of mood in verbal stem formation is discussed in the following subsection.

### 4.1.3 Derivation versus Inflection

Nearly every treatment of morphology in language inevitably touches on the distinction between derivation and inflection, which Aronoff (1976: 2) rightly calls "delicate". Although the difference between derivation and inflection is similar cross-linguistically, I agree with Dixon (2010: 220-21) that the difference is not universal and absolute: the difference must be established on a case-by-case basis. Aronoff (1976: 2) and Dixon (2010: 218–21) give helpful indications of possible ways to distinguish derivation and inflection, but Aronoff's (*loc. cit.*) assertion that "Derivational morphology is thus restricted to the domain of *lexical category*" is too strong. Better is the formulation Booij (2007: 51): "The **basic function** [emphasis mine – RS] of derivational processes is to enable the language user to make new lexemes." If one wishes to accept this definition outright, however, then an strict definition of "lexeme" is needed, which I am not prepared to offer. When a morphological process clearly serves to form a word that is semantically distinct from or changes the syntactic category of the word to which it applies, it is almost surely derivational. Beyond this rule of thumb, the matter is more fraught.

A serious concern is whether, in fact, as Aronoff (1976: 2) claims, "derivational morphology is not paradigmatic". To return to an earlier example, what is the synchronic relationship between *pináṣṭi* 'crushes' (3.sg.pres.act.ind.), *ápikṣan* 'crushed' (3.pl.aor.act.ind.), and *pipéṣa* 'has crushed' (3.sg.perf.act.ind.)? Do the stems exist in some sort of derivational paradigm? In earlier research, I have presented possible evidence of a derivational paradigm for reduplicated verbal stems in Vedic Sanskrit, in which the perfect stem appears to serve as a base for forming other reduplicated verbal stems; this argument was grounded in statistical evidence that verbal roots in Vedic are much more likely to attest other stems formed with reduplication if they attest a reduplicated perfect (Sandell 2011a). Similarly, the Indo-European "Caland"-System gives the appearance of a sort of derivational paradigm, in which the existence of a primary derivative of some type may permit the derivation of another primary derivative (Caland 1892, Nussbaum 1976). Hence, the existence of paradigmatic relations between word forms does not constitute *prima facie* evidence of inflection. What Aronoff may mean, more precisely, is that derivational processes are never obligatory, and the existence of some derivational form cannot necessarily be predicted on the basis of another, whereas, excepting instances of paradigm gaps, the existence of other inflectional forms is expected

[3]Note that, confusingly, the term "present" is traditionally used to mean both the imperfective aspectual stem and present tense, while "imperfect" then refers to the past tense usage of the "present" (imperfective) stem.

and predictable.[4]

Morphological processes that serve to mark "grammatical" or "syntactic" functions alone (Aronoff 1976: *loc. cit.*) are not necessarily inflectional. Dixon (2010: 221) observes that "in Jarawara, there are many grammatical systems realized through suffixes – mood, tense, evidentiality, modality, negation, etc. – but all are optional. It is not at all helpful to try to categorize these as derivational or inflectional."[5] The criterion of "optionality" to which Dixon alludes here is very useful, I think, for languages like Greek and Sanskrit. Certain grammatical information is obligatorily marked on every noun and verb: case and number for nouns, person, number and voice, for verbs. Although every verb in those two languages also has an aspect and a mood, and every noun has a gender, specific markers of those categories are not obligatory.[6] Nominal gender is uncontroversially a feature of suffixes that seem derivational. That aspectual stems are derivational in nature is apparent from their derivational relations: for example, the root aorist stem Grk. στη- [stɛ:-] 'stood up' has a corresponding present stem ἵστη- [hístɛ:-]. Meanwhile, the place of the modal affixes for subjunctive and optative is difficult to establish. Although a modal affix is never obligatory, since the indicative is unmarked and the imperative is expressed in a special set of obligatory endings, every verb has a modal reading based on formal features alone, unlike aspect, the reading of which depends on both formal and lexical factors.

The conclusion concerning derivation and inflection for the purposes of this study is then, essentially uncontroversial: the markers of case/number and person/number/voice/(tense) in nouns and verbs, respectively, constitute inflection; other suffixes preceding the endings, with perhaps the exception of subjunctive and optative markers in the verb, are derivational. I will therefore leave the subjunctive and optative aside, but regard any other suffix as part of derivation. Although compounding processes are derivational as well, I will limit my study to purely affixal morphology. For practical purposes then, all verbal suffixes concerning aspect and valency, and nominal stem-forming suffixes, are presumed to fall within the bounds of Sanskrit and Greek derivational morphology.

## 4.2 The Corpora

In principle, a corpus ought to be a representative sample of language, which entails the inclusion of speech and/or writing in a wide variety of styles, registers, and genres, from many different authors or speakers. In the study of corpus languages, however, to fulfill this objective is not a realistic possibility; the genres of the texts are very limited, often to essentially single type of text, if not to even a single text altogether. As Baayen (1989: 30) observes, however, "theoretically no text corpus is a representative random sample of a given language L

[4] In addition, even inflectional processes of low utility typically show very high quantitative degrees of productivity. See further discussion at 8.2.

[5] See further Dixon 2004: chs. 6 & 7.

[6] Where exactly tense falls is somewhat more difficult to say, but in both Greek and Sanskrit, obligatory personal endings do normally encode a distinction between past and non-past, but forms having endings that normally occur with past tense forms may also express present or future tense in some circumstances. The correct distinction, then, is probably between non-past and unmarked tense.

…a corpus is at best exemplary of the population of utterances." That is to say, to construct an ideal corpus is not a living possibility in any case. If one wishes to carry out corpus-based studies on corpus languages, one must simply accept the limitations of those corpora, and try to compensate for their weaknesses when reporting results.

From the point of view of the study of Indo-European morphology, the choice of languages and corpora seems straightforward: the best choices are the two Homeric epics (the *Iliad* and the *Odyssey*), for Greek, and the *Ṛgveda*, for Vedic Sanskrit. Looking to other old Indo-European languages, no other language of comparable antiquity is attested in comparable quantity and philological reliability. The Hittite corpus is at least as old or older than the *Ṛgveda*, but the overall size of the corpus is smaller than either Homer or the *Ṛgveda*, and the often broken condition of the texts, and their often uncertain restoration, presents a host of problems. Thus, while the Hittite corpus is in one sense more philologically secure than the Greek or Vedic, since it is not mediated by millennia of transmission, to use it, at this time, for large-scale corpus research would have practical difficulties. Old Latin presents a corpus of adequate size and condition for profitable study, but the date of the language ($3^{rd} - 2^{nd}$ c. BCE) makes it less immediately relevant for the discussion Indo-European morphology. Some of the work in the case studies will represent the first step towards a more complete quantitative characterization of Indo-European morphological processes.

Another important concern is the availability of electronic text editions that can facilitate the retrieval of forms. Simple .txt files of the RV are available at:
`http://gretil.sub.uni-goettingen.de/`. Text retrieval from $20^{th}$ century editions of Homer is readily available in a variety of guises, some of which are discussed further in Chapter 5. In addition, the amount of printed material, in the form of concordances, word-indexes, dictionaries, and grammatical monographs on Homer and the RV ensures that obtaining accurate type lists and token counts of morphological categories as they appear in those two texts is not prohibitively onerous.

The only serious worry that remains is the size of the corpora themselves. In comparison to corpora of modern languages on the order of at least several (hundred) million tokens, both the RV and Homer are tiny, each containing under 200,000 tokens. Nevertheless, Baayen's work with a relatively small corpus of Dutch (the Eindhoven corpus, c. 600,000 words), Březina's (2005) study using a 450,000 word corpus of Early Modern English, and the small studies carried out using the text of *Moby-Dick* in Chapter 2 seem to have produced reliable results. In part, the function of the case study in Chapter 5 is to demonstrate that the size of the Homeric corpus (which is slightly smaller than the Vedic corpus) is generally adequate to draw linguistically interesting and accurate conclusions.

### 4.2.1 The *Ṛgveda*

The *Ṛgveda* is a collection of principally religious praise and ritual poetry, consisting of approximately 40000 lines of verse, divided over 1028 hymns in a variety of syllable-counting meters, collected into ten books (*maṇḍalas*). Electronic word counts of the saṃhitāpāṭha

and padapāṭha texts[7] count 159430 and 166672 tokens, respectively, and the latter number should be closer to an accurate estimate; the padapāṭha text counts 30995 distinct surface word types. It is assuredly the oldest Sanskrit text, the composition of its contents likely spanning approximately two centuries during the later half of the second millennium BCE (thus ∼ 1400–1000 BCE); cf. Jamison and Brereton 2014: 5. Books II–VII, the so-called "Family Books" because their authorship is attributed to a single family group, as well as many hymns of Book IX (all dedicated to the plant *soma*), which were extracted from the family collections, are chronologically older, while most of Book X is chronologically later.

The exact time and conditions under which the Ṛgvedic hymns were organized into a collection resembling their present form are uncertain; Witzel (2003) suggests that this process began with the emergence of proto-states among the Vedic tribes somewhat before 1000 BCE. The hymns were then transmitted orally in a number of recitational schools; while we know that numerous recensions of the RV (at least five, probably) associated with Vedic schools existed, only the recension of the Śākala school has come down to the present day, with the oldest manuscripts dating to the 14$^{th}$ c. CE. There is no evidence that the RV or any other orally transmitted Vedic texts were fixed in writing prior to ∼ 1000 CE. Despite this long history, as Jamison and Brereton (2014: 17–8) say, "The Ṛgvedic tradition has preserved a very ancient literature with extraordinary fidelity, with no grammatical or lexical modernization or adjustment of contents to later conceptual conditions." The text sometimes betrays phonological and morphological features that are typical of Middle Indic languages or later Sanskrit, rather than Vedic; these features are not intrusions or later alterations, but rather seem to reflect a lower register or "home" language of the poets, in contrast to the usual high-register hieratic language of most of the text.

For further details on the nature, genesis, and literary features of the Ṛgvedic text, I refer the reader to the Introduction of Jamison and Brereton 2014 and citations therein.

### 4.2.2 The Homeric Epics

As a genre of Greek literature, *epic* refers to a set of fairly lengthy metrical texts that concern the deeds of traditional heroes and gods, as well as what one might consider folk wisdom. The two epic poems traditionally attributed to Homer, the *Iliad* and the *Odyssey*, are the principal exponents of the genre, along with works of Hesiod, and later imitators active during the Hellenistic period (such as Apollonius of Rhodes). Altogether, the *Iliad* and *Odyssey* consist of approximately 199000 tokens (from one electronic text file, I count slightly less, while the Perseus Project (`http://www.perseus.tufts.edu/`) counts slightly more), with the former being considerably longer than the latter. Both poems are composed entirely in lines of dactylic hexameter, that is, each line consists of six feet of the form — ⏑⏑, the sixth foot ending with an anceps. For an introduction to further features typical of the hexameter, including the division of the line into cola and types of metrical licenses admitted, see West 1982: 35–9.

[7] I made some modifications to the padapāṭha text, principally uniting the so-called pada case endings such as inst.pl. *-bhiḥ* with their stems.

Like the Vedic lyric poetry transmitted in the *Ṛgveda*, the Homeric epics were a type of traditional oral poetry, though unlike the poetry of the *Ṛgveda*, the content of the poems was never consciously and rigorously maintained in a single form, until it was committed to writing. That is to say, the Greek poets of the late second and early first millennia BCE learned a set of traditional stories and songs, including a language specific to the telling of those stories, but recomposed the precise form at each retelling; this type of oral poetry and its transmission agrees substantially with the situation documented among Serbo-Croatian epic poets during the 20$^{th}$ century (cf. Lord 1960). The two poems are usually believed to have assumed something resembling their present form during the 8$^{th}$ c. BCE, and are linguistically consistent with such a dating. Widely accepted scholarly speculation suggests that the poems were standardized and perhaps committed to writing at Athens during the Peristratid dictatorship of the 6$^{th}$ c. BCE. Substantial philological work on the poems, including the introduction of spaces between words and prosodic notation, were introduced by scholars, working mainly at the Library of Alexandria, from the Hellenistic period forward, and later at Byzantium. The principal manuscripts that form the basis of modern editions were produced in Byzantium (e.g., the Venetus A manuscript, datable to the 10$^{th}$ c. CE). For further details on the history and transmission of the Homeric texts, a good introductory reference is Reynolds and Wilson 1991: 5–16; more detailed and specialized overviews of the issues, from different viewpoints, include Haslam 1997 [2011], West 1998: VI–XVI, and Nagy 2004.

The principal language of the poems is a form of East Ionic Greek, consistent with the varieties of Greek spoken on the Western coast of Anatolia during the first millennium BCE. However, at least at the level of phonology and morphology, some material belonging to "Aeolic" dialects (as traditionally construed) and Attic appears, though the variation that they contribute, in most cases, should wash out, I believe, in a statistical analysis. Nevertheless, the language of Homer remains a *Kunstsprache*, containing such dialect mixture in part as a device that afforded singers a greater number of variant forms that could be employed to satisfy metrical exigencies.

### 4.2.3 Designing Case Studies

The purpose of this dissertation is not to provide a catalogue of measurements on morphological productivity in Vedic and Greek, but rather to demonstrate how such measurements can play a role in historical linguistic research. The judicious selection of case studies best serves that end. In particular, the study of cases of morphological competition, i.e., where a coherent functional category is potentially expressed by more than one type of formation, is important: morphological competition can best illustrate changes in productivity that depend upon system-internal dynamics. Indeed, Gaeta and Ricca (2006) claim that the study of morphological productivity is really only interesting or meaningful where competition exists. Chapter 5 presents a clear case study concerning morphological competition, wherein one type is replacing other types.

To build a case study entails not only the collection of all the word types belonging to a particular morphological formation, but also all of their tokens. This first step allows for the

calculation of productivity measures. I wish, however, to delve deeper, in order to bring out the factors that affect the (non)-productivity of a formation. In general, whenever possible, to produce case studies of cognate categories in Vedic and Greek seems sensible. Hence, Chapters 5 and 6 both look into the cognate aorist forms of the two languages, while Chapter 7 looks at the productivity and accentuation of cognate nominal categories.

### 4.2.4 Data Collection

The primary obstacle to the efficient study of morphological productivity in Greek and Vedic, is the absence of morphologically analyzed corpora, comparable to the CELEX corpus for English, Dutch, and German. Since the process of constructing such corpora would be too time-consuming, existing philological materials will be sufficient for the purpose of collecting necessary token counts and type lists. Nevertheless, Evert and Lüdeling (2001) mention that a considerable amount of manual pre- and postprocessing was needed in the extraction of data from their German corpora. At a surface level, certain processes like ablaut cannot be extracted effectively at all. In the absence of deeply tagged texts, the manual extraction and validation of types is necessary, and allows for quality control of the data, despite being time-consuming.

For Homeric Greek, the work of Risch (1974) on word-formation in both the noun and the verb is indispensable. In many cases, Risch provides complete type lists, and his linguistic and philological commentary greatly facilitates the treatment and evaluation of individual datapoints. The electronic search capabilities provided by WordHoard (2004–2011) and the *Thesaurus Linguae Graecae* (TLG) makes the process of finding tokens relatively easy. The only considerable disadvantage is that, because all derivational stems associated with a verb are traditionally arranged under a single lemma (a sometimes hypothetical 1.sg.pres.act.ind. form), one cannot neatly pull together lists of forms derived through the same process. Concordances (e.g., Tebben 1994) can also supplement electronic searches when necessary. WordHoard (2004–2011) and the *TLG* complement one another, in that WordHoard offers the possibility of search by morphosyntactic tags, while the *TLG* has compatibility with regular expressions.

The resources for work on Vedic are more disparate. I possess electronic text files of both the saṃhitāpāṭha and padapāṭha versions of the *Ṛgveda*, following the edition of Aufrecht (1877), as well as the "metrically restored" version of van Nooten and Holland (1995) from which I can search for simple patterns or search using regular expressions to find datapoints.[8] No sort of morphological index as complete as Risch 1974 exists for the *Ṛgveda*, but the dictionary of Grassmann (1872 [1976]) includes an invaluable stem index. Whitney 1885 [1963] and Avery 1880 offer accessible materials for systematically the systematic search verbal stems, and specific monographs on verbal categories (e.g., Narten 1964 on sigmatic aorists, Cardona 1960 on thematic aorists) are of evident utility as well.

[8] The "metrical restoration" is the undoing of many persistent surface-level phonological phenomena (e.g., vowel elision, semivowel formation) that characterize later Sanskrit and that became part of the RV's transmission, so as to render many lines metrically correct.

Almost inevitably, this manual process of data collection will result in some oversights and errors. However, where the category under study is sufficiently large, I expect the number of such errors to be insignificant statistically. In the long term, it is to be hoped that a collaborative effort between technically competent Indo-European linguists and philologists will be able to produce accessible tagged corpora of many old texts. See the closing Summary and Conclusions for more on this point.

## 4.3 Issues in the Corpus-Based Study of Morphology

### 4.3.1 What Do We Count?

Perhaps surprisingly, exactly what forms should be counted as words, let alone which words count as members (or not) of a given category, is a strikingly difficult problem. Even the basic choice of choosing how to count the types of the category is not totally uncontroversial – are the types each distinct inflectional form (e.g., Skt. *ūtíḥ* 'aid (nom.sg.), *ūtím* 'aid (acc.sg.), *ūtáye* 'aid (dat.sg.)', etc.), or only distinct lemmata (*ūtí-* 'aid', *kṣití-* 'ploughed land; people')? How to count other potentially relevant frequencies is even tricker; Ford et al. (2003: 97–8) describe four different measures of morphological family size (each distinct derivational form; each distinct derivational form and compounds; each distinct inflectional form of each distinct derivational form; the preceding plus compounds).[9]

The most pressing question in the counting of word types is the matter of inner-cycle derivation: should words that contain multiple derivational processes count as instances of all the processes involved, or count only as an instance of the "last" or "outermost" derivation to apply? For instance, in a form such as the sigmatic aorist middle participle ἐρυσσάμενο- [erussámeno-] 'drawing out', one could count only the outermost derivation, the middle participle suffix /-meno-/, or as both that and the sigmatic aorist. Many quantitative studies of morphological productivity (e.g., Plag 1999: 29) indeed operate with the outermost cycle alone; for languages without the added complication of inflection or substantial (orthographic) allomorphy, finding the outermost prefix or suffix is easiest from a practical point of view. Baayen (1989) also notes that counting all derivational cycles removes the assumption of independence for each process, which diminishes the reliability of some statistical tests that one might wish to use to evaluate quantitative differences between processes.

Gaeta and Ricca (2006: 79), however, rightly point out that the decision to count only outermost derivational processes is both theoretically problematic in its application to individual forms, and makes little sense from the psycholinguistic point of view. As an example of the first issue, consider another Greek aorist participle, ἐπευξάμενο- [epeuksámeno-] 'praying to': while the traditional lemmatization of the form as belonging to a verb ἐπεύχομαι 'pray to' implies that the derivation of the aorist aspect and middle valency are outside the the prefix /epi-/, whether the prefix is considered outermost or innermost, i.e., whether the bracketing is [ [ [ epi [euk] ] sa] meno- ] or [ epi [ [ [euk] sa] meno- ] ], produces no evi-

[9] Ford et al. (2003: 101) also describe some interesting procedures used to eliminate collinearity in a number of frequency variables.

dent difference in interpretation. Cases such as this one create conundrums if one seeks to consistently count only outermost derivations. From a practical point of view, not to count inner cycle derivations may inappropriately inflate the number of hapax legomena, by failing to consider potentially accessible inner derivations; for instance, all sigmatic aorists in Homer that attest only one finite form would count as hapax legomena, since the further derived participles and infinitives would not go towards the token count of the aorist stem. At the same time, other potential hapax legomena embedded within further derivation might be missed. The same consideration applies to morphologically complex forms embedded in compounds: πολυμῆτι- [polumêti-] 'of many wiles' can count not only towards the token count of that bahuvrīhi compound itself, but towards the token count of the *u*-stem adjective πολύ- [polu-] and *si-/ti*-stem noun μῆτι- [mêti-].

Gaeta and Ricca (2006), operating with a corpus of Italian, found that counting all derivations present or only the outermost cycle did not affect the relative productivity rankings, though, by adding considerably more tokens to the count of each process, reduces the absolute $\mathscr{P}$ measures. For the relatively small corpora that constitute the basis of my studies, I believe that the loss of psycholinguistically justified data that would result from counting only outermost cycles is of greater concern than the weakening of independence assumptions.[10] Consequently, I will try throughout to count embedded derivations when collecting frequency data for a morphological category, insofar as practically feasible.

### 4.3.2 Base : Derivative Relative Frequency

One specific circumstance under which the counting of inner derivation might justifiably be ignored, in accordance with phenomena discussed under 3.2 and 3.4, is when a derivative exhibits a token frequency that is greater than its base. The presumption in such a case would be that the more frequent derivative would be less parsable, and consequently the embedded derivational process less accessible to the speaker. Nevertheless, if forms belonging to categories with high $\mathscr{P}$ are precisely more parsable, then the outer derivation of a form might remain transparent, and the embedded derivation accessible, despite a higher derivative frequency – while base : derivative frequency is relevant to the mode of access for individual lexemes, its effect is tempered by the overall productivity of a process.

As an example, consider the derivational prefixes in Greek that add semantic specification or alter the Aktionsart of a base verb, e.g., πνέω [pnéɔː] 'breathe, blow' and ἀναπνέω [anapnéɔː] 'blow through'. Although, in Homer, prefixed verbal forms not uncommonly have a token frequency greater than their base, virtually all of these prefixes are quantifiable as highly productive, with $\mathscr{P}$ between $0.08$ and $0.1$. There is, therefore, good reason to add the frequencies of the prefixed forms to the base forms, or even set up a base verb that does not occur in the corpus (e.g., Homer attests root aorists ἀποπλω- [apoplɔː-] 'swim away' and

[10] That is to say, I believe that under present circumstances, loss of reliability from small sample size is of greater concern than loss of reliability from lack of independence. Baayen et al. (2003: 474) conclude that "studies using frequency norms based on small corpora are especially prone to sampling error" (though they do not define what "small" is, precisely), and such sampling error is best avoided by counting as many elements as possible.

ἐπιπλω- [epiplɔː-] ‘swim upon’, but the not simplex πλω-* [plɔː-]*) and add the frequencies of the derivatives.[11] The simplest procedure, overall, is thus always to add inner derivation frequencies to the frequency of a base, though not to add those tokens in the case of an unproductive outer derivation is a justifiable choice that might better approximate the psycholinguistic reality.

### 4.3.3 The Reliability of Corpus Frequencies

That the lexical frequencies extractable from corpora provide a realistic representation of lexical frequencies in actual usage can and has been criticized on several grounds. One such concern is simply the representativeness of the sample corpus for the language as a whole; this problem impacts any research involving corpus-linguistic data, and can be resolved only through the construction of corpora that would be more “representative”. For the purposes of this study, history has wholly eliminated the option of building or selecting better corpora – we cannot improve upon or expand the corpora without encountering the problem of including material from too many different chronological strata.

More fundamentally worrisome is the claim of Gernsbacher (1984) that word frequencies, when compared across different corpora of the same language, might exhibit “regression towards the mean”. Regression towards the mean is typical of phenomena that conform to the normal (or Gaussian) distribution: an item that stands below the true population mean in one sample would likely have a greater value in another sample, while an item that stands above the true population mean would likely have a lesser value in another sample. If word frequencies were indeed subject to regression towards the mean, then the distribution of very high and very low frequency lexemes could be substantially different from sample to sample (corpus to corpus). Consequently, both the validity of psycholinguistic work centered around a fundamentally real difference in lexical frequency where the frequency data derives from one corpus would be jeopardized.[12] Similarly, a corpus-based measure of productivity like $\mathscr{P}$ could take on wildly different values for one category or another when using different corpora, because the set of hapax legomena and category-conditioned token frequencies might be radically different as lexical frequencies distributed themselves differently.

Thankfully, Baayen et al. (2003) have demonstrated that word frequencies, both theoretically and empirically are not subject to regression towards the mean. Theoretically, of course, as we have seen in Chapter 2, word frequencies of natural languages, in even in relatively small samples, simply are not normally distributed; they follow the Zipfian or some mathematically similar distribution. Baayen (2001: 57–63) demonstrates that the Good-Turing estimate (Good 1953) can be used to obtain an estimate of the expected number of tokens of an item in a sample given the actually sampled number of tokens. For high-frequency items in a corpus of one million words, the difference between the expected value and the sample value is vanishingly small, while for a hapax legomenon, the expected value is $0.68$

[11] See 5.2 for further details on this particular issue.

[12] Gernsbacher (1984)’s concern was motivated by the fact that a substantial amount of psycholinguistic work on lexical processing used frequency data derived from the Brown corpus (Kučera and Francis 1967).

(which means, in effect, that the word is more likely to occur than not). For smaller corpora, the differences between expected and sample values will be somewhat greater, but not worrisomely so. Empirically, Baayen et al. show that the frequencies of specific lexemes between 20 one-million word subcorpora of the British National Corpus correlate strongly with one another, and that likewise the frequencies across the 18-million word CELEX corpus and an 18-million word corpus built from samples of the British National Corpus strongly correlate. There is no real evidence of regression towards the mean in word frequency distributions, and hence, the distribution in one sample corpus is just as reliable as another, in this regard.

However, language corpora do face a genuine problem of sampling error in the form of *underdispersion*: words do not have a smooth distribution within one corpus or between corpora. Hence, some words that may occur several times in one corpus may not occur whatsoever in another. This effect is due to topicality: the content of a corpus partly determines what lexemes will be used therein. Given the content of the Homeric epics and the *Ṛgveda*, one might reasonably think that the frequency of occurrence for the names of gods and heroes, or items relating to battle or the *soma* ritual are overrepresented. The easiest cure for underdispersion is simply to obtain a larger corpus containing more texts concerned with a greater variety of topics. Gries (2008) develops a measure for determining the degree of underdispersion of a lexeme within a given corpus, and his applications show that, unsurprisingly, high-frequency lexemes generally are less underdispersed, and low-frequency lexemes, with hapax legomena being a limiting case of maximal underdispersion, are more underdispersed. The highest-frequency lexemes in a corpus thus probably do accurately reflect the real proportion of tokens in the language as a whole that those lexemes constitute.

For the calculation of productivity measures, underdispersion does not pose an immediate problem, since any given rare or novel lexeme could happen to appear in the corpus. Underdispersion is only a concern insofar as the occurrence of a lexeme belonging to one category might induce the repeated reuse of that lexeme, or trigger the use of other lexemes belonging to that same category. In the former case, the increase in tokens and decrease in HL will give the appearance of less productivity; in the latter, the occurrence of more types and HL will give the appearance of greater productivity. For instance, the occurrence of three root aorist subjunctives, all hapax legomena root aorists, in a single line in the RV (2.30.7a; cf. the forms *tandrat*, *tamat*, and *śramat* in Table 6.12) may reflect just such an instance of creative triggering. For all of these potential issues, however, none of the existing work on the corpus-based measurement of productivity has raised the possibility that underdispersion might create serious problems of sampling error, beyond the potential for sampling error introduced by small corpus size. At present, I think that it is appropriate to acknowledge underdispersion as a potential issue, but for want of well-developed means of adjusting empirical frequencies to compensate for underdispersion, I will simply accept the raw word frequencies in my corpora. In part, the studies in Chapters 5 and 6 serve to demonstrate that, practically speaking, underdispersion does not create notable problems for the measurement of productivity in Homer or the RV.

Part II

# Case Studies

# CHAPTER 5

# The Aorist in Ancient Greek

This case study concerns the productivity of aorist formations in Ancient Greek, taking the two Homeric epics, the *Iliad* and the *Odyssey*, as the corpus for the purposes of data analysis. The measurement of productivity using $\mathscr{P}$ and $\mathscr{P}*$ happily points to the productivity of sigmatic aorist formations, just as the later continuing history of the Greek aorist would suggest. These results are a good thing: Baayen and Lieber (1991) point out that a productivity measure should concord with native intuitions about productivity, so if the measures concord with the intuitions of philologists, we thankfully do not need to claim that either the measures or the intuitions are basically flawed. Analysis of the factors underlying the productivity of the sigmatic aorist will reveal that, if one proceeds from the assumption that the present stem serves as a base of derivation for the aorist stem, sigmatic aorists are largely transparent (being secondary or tertiary derivatives), whereas root and thematic aorists are largely opaque (primary derivatives). This fact suggests that speakers rely on existing transparent derivational patterns for producing new forms. Further examination of the same categories in the Greek of the New Testament ($\sim$ 800 years later) finds little change in the productivity *per se* of the aorist categories, but that the number of types belonging to moribund categories is indeed fewer, and some aorist stems attested in both Homer and the NT have changed category, going over to the productive category. By and large, comparison of the same aorist formations from the later New Testament indicates a stable maintenance of productivity among sigmatic aorists (with the extension of a few small derivational patterns therein), and gradual attrition of root and thematic aorists.

## 5.1 Morphological Characteristics and Problematization

The structure of the verbal system in Ancient Greek is substantially similar to the core features reconstructed for PIE. Most salient for the present purpose is the essential division of verbal stems into imperfective ("present") and perfective ("aorist") aspectual stems. The traditional grammatical description of the Ancient Greek verb recognizes three basic types of aorist stem for the active and middle voices:[1]

1. ROOT or ATHEMATIC aorist: the inflectional ending that marks person, number, and voice attaches directly to the verbal root; the root of some such aorists (namely, those taking the athematic *-mi* class of personal endings in their present stems, rather than

[1] Certain intransitive, and all passive, aorists have other particular morphological formants: the suffixes -η- [-ɛ:-] or -ϑη- [-tʰɛ:-]. Aorists built with these formants are excluded from further consideration.

| | |
|---|---|
| | ATHEMATIC |
| A01 | athematic, asigmatic (e.g., ἔβην [ébɛːn] 'I went', ἔγνων [égnɔːn] 'I recognized') |
| A02 | athematic, asigmatic + suffix -κ- in the sg.act.ind. (e.g., ἔδωκα [édɔːka] 'I gave'), with alpha-thematic inflection in the κ-forms |
| A03 | alpha-thematic (e.g., ἔχεα [ékʰea] 'I poured', ἤνεικα [éːneika] 'I reached, arrived') |
| | THEMATIC |
| A04 | thematic, asigmatic, with zero grade of the root (e.g., σχεῖν [skʰêːn] 'to have taken') |
| A05 | thematic, asigmatic, with other root vocalism (e.g., ἔτεκον [étekon] 'I begot', ἔθορον [étʰoron] 'I leapt') |
| A06 | thematic, asigmatic, with reduplication (e.g., ἔπεφνον [épepʰnon] 'I killed') |
| | SIGMATIC |
| A07 | sigmatic (e.g., ἔδειξα [édeiksa] 'I showed', ἔστειλα [éste:la] 'I sent'); transitive-causative to intransitives (e.g., ἔβησα [ébɛːsa] 'I made go') |
| A08 | -ησα [-ɛːsa] (e.g., ἐδέησα [edéɛːsa] 'I lacked') |
| A09 | sigmatic-thematic ["aoristus mixtus"] (e.g., ἵξον [híkson] 'they arrived') |

Table 5.1: Classification of Greek Aorist Types

the thematic -*ō* class) shows ablaut, with Indo-European full-grade *[e] in the active singular and the subjunctive mood, but Indo-European zero grade elsewhere. Some varieties of root aorist show a linking thematic vowel -α- [-a-] between root and inflection, and are thus termed "alpha-thematic".

2. THEMATIC aorist: an ablauting suffix -ε/ο- [-e/o-] follows the verbal root.

3. SIGMATIC aorist: a suffix -σ(α)- [-s(a)-] follows the verbal root.

For the general formation of these types, see Agazzi and Vilardo 2002: Ch. 16, or Schwyzer 1938 [1953]: 739–56 for fuller specific details, and especially for the current study, Risch 1974: 233–50. On the Indo-European origins and reconstruction of these formations, see Rix 1976 [1992] and Rix et al. 2001. Each of these three aorist types may be subdivided into several subtypes, for which I follow the classification scheme of van de Laar (2000: xv), given in Table 5.1. Examples there are cited in either the 1.sg.aor.act.ind. or aor.inf.act.[2]

Since the categories listed in Table 5.1 will ultimately be examined at the level of ATHEMATIC, THEMATIC, and SIGMATIC, the precise classification of some ambiguous forms internal to those groups is not of great consequence. For instance, the stem εὑρο- [heuro-] 'found' is evidently thematic, but whether the form reflects a historically reduplicated form */u̯e-u̯rh$_1$-e/o-/, or simple */u̯erh$_1$-e/o-/ is not entirely certain (see discussion in Peters 1980: 22 ff. and Beckwith 1994), and even if reduplicated historically, whether the form is synchronically

[2]The prefix ἐ- [e-] occurring in the forms there is the so-called 'augment', which encodes past tense, as does likewise its Sanskrit cognate *a-*. In this chapter, when speaking of aorist stems, and not specific forms, I will always omit the augment.

parsable as such is uncertain. In that particular case, I have labeled εὑρο- [heuro-] as Ao5, but to treat the form as Ao6 instead would have no real consequence for this study.

Synchronically, I have separated class Ao7 into two types, which I label as Ao7 and Ao7b. Ao7 shows an overt -σ- [-s-] marker, whereas Ao7b is limited to roots ending in liquids (-ρ- [-r-] and -λ- [-l-]) or -ν- [-n-], where the *[-s-] was lost and the preceding consonant moved into the onset of the following syllable, thus triggering compensatory lengthening of the preceding vowel, e.g., *[stel.sa-] > [steː.la-].[3] One might then reasonably regard aorists with [-VːR-] as allomorphs of /-s-/, i.e., [-VːR-] ← /-VR-s-/. However, sometimes both a type Ao7 and a type Ao7b to the same root occur in Homer, for instance, both Ao7b κειρα- [keːra-] and Ao7 κερσα- [kersa-] as aorists to κείρω [keírɔː] 'cut' are found; sigmatic aorists that permit an [s] following a sonorant are rare to non-existent outside of Homer.[4] Similarly, to find completely different and competing aorist formations to the same root is hardly uncommon in Homer, e.g., Ao7 τευξα- [teuksa-] and Ao6 τετυκο- [tetuko-] to τευχω [teukʰɔː] 'make, produce, prepare', or Ao4 φανο- [pʰano-] and Ao7b φηνα- [pʰɛːna-] to φαίνω [pʰainɔː] 'shine'. Such co-occurrence of multiple aorists is a relative rarity the New Testament, in contrast.

The fact that fundamentally different formations occur to the same root is interesting, from the point of view of morphological competition, given that, in large part, all of these aorist formations fulfill the same function: they provide the perfective past stem within a verbal paradigm. The only notable semantic distribution for any type is for roots with intransitive meaning, e.g., βη- [bɛː-] 'go, come' or στη- [stɛː-] 'stand', to build corresponding causatives with a sigmatic aorist, thus βησα- [bɛːsa-] and στησα- [stɛːsa-]. Apart from these cases, the problem of morphological competition here is roughly similar to the competition between English abstract nominalizing suffixes *-ness* and *-ity*: the two affixes fulfill similar conceptual/semantic functions, but have different degrees of productivity and distributional restrictions (cf. *inter alia*, Aronoff 1976: 37–45, Baayen 1992: 133–8). Thus, for the Greek aorists, some distributional restrictions likely exist, but the co-occurrence of more than one type could be the result of the encroachment of productive types.

## 5.2 Data Collection and Preparation

In order to discuss and measure the productivity of the aorist formations, using either Baayen's corpus-based or MGL measures, the obvious prerequisite is a database of all aorist stems that

[3] The compensatory lengthening surrounding these sigmatic aorists would be an instance of "double flop" compensatory lengthening as formalized in Hayes 1989. It is interesting to note that both Ionic and Attic dialects exhibit this lengthening in aorists and futures built with an [s]-suffix, but only (East) Ionic shows "double flop" lengthening from the historically later loss of [w] (e.g., *[kal.wo-] 'good' > East Ionic [kaːlo-] but Attic [kalo-]). The implication is that the conditions in the phonological grammar to admit of "double flop" compensatory lengthening must have been present in Proto-Attic-Ionic; the grammar of Attic then must be innovative on this point with respect to its last shared ancestor with Ionic.

[4] In fact, the history surrounding the treatment of word-internal [-ls-] and [-rs-] across the Greek dialects, and in aorist forms in particular, is very complex. The issue is at most tangential to the concerns of the present chapter, and I therefore refer the reader to the overview in Lejeune 1972: 124–6 and detailed survey in Forbes 1958.

occur in the Homeric epics and their frequencies. The information in that database would serve to provide the following necessary datapoints:

1. the number of types ($V$) belonging to each (sub)type of formation.
2. the number of tokens ($N$) belonging to each (sub)type of formation.
3. the number of hapax legomena ($n_1$) belonging to each (sub)type of formation.

Knowledge of these frequencies permits the calculation of $\mathscr{P}$, $\mathscr{P}*$, $\mathscr{I}$, and the $n_1/V$ ratio, following the procedures described in 2.2. An ordered list of stems also facilitates data input for an MGL simulation.

The basis for building my database is the WordHoard 2004–2011 software and text repository. WordHoard provides access to tagged corpora of both Homeric epics, based on the work of the Chicago Homer (`http://homer.library.northwestern.edu/`). The text of the *Iliad* used for this corpus is Monro 1902; the text of the *Odyssey* is taken from Murray 1919.[5] Each word in these corpora is tagged with a lemma, part of speech, morphosyntactic features, and other data. For morphosyntactic tags, a sequence of numbers corresponds to a bundle of morphosyntactic features, e.g., 1210013, in the tagging scheme employed, indicates a 3.sg.pres.act.subj. Thus, one can, using the WordHoard program, search for all forms given a particular morphosyntactic tag, and sort the display of forms by lemma and frequency. Through such queries to the WordHoard program, I called up, one combination of morphosyntactic features at a time, all aorist active and middle forms that occur in the *Iliad* and *Odyssey*. To a great extent, my results are then dependent upon the work of tagging done by others: if a form was not tagged as an aorist of any sort, I have not seen it; conversely, I could eliminate a handful of forms that I judged to be erroneously tagged as aorists.[6]

For the purposes of organizing the data, I prepared a spreadsheet, with columns for the lemma (i.e., the dictionary headword under which the aorist stem would be listed in a work like the Liddell Scott Jones Lexicon), the aorist stem, the aorist type following the modified van de Laar classification scheme described above, and columns for every combination of morphosyntactic features relevant to aorists (1.sg.act.ind, inf.mid, gen.pl.mid.part., etc.). A given row then contains the form of a lemma, the form of an aorist stem, the classification number, and the number of tokens attested to a given inflectional form that the stem shows. Table 5.2 illustrates this arrangement.

[5] The editor of WordHoard 2004–2011 adds that readings from van Thiel 1991 (Odyssey) and van Thiel 1996 (Iliad) have sometimes been followed. In any case, the editorial choices are likely to have little or no impact on the sort of derivational morphology relevant to the present study. In principle, because the assessment of productivity depends heavily on hapax legomena, it is mainly those forms that should be given closer examination. Furthermore, since different aorist formations are rarely, if ever, isometrical, the possibility that an editor would have the option to select a variant form is highly unlikely. Indeed, the main possible source for isometricity among aorists would be precisely in forms like A07b κειρα- [ke:ra-] : A07 κερσα- [kersa-] mentioned above (though in this particular case, the choice of stems seems to depend on diathesis; cf Chantraine 1958: 173 or Forbes 1958: 268), for which the fact that both types remain sigmatic aorists means that no effect on the overall sigmatic aorist category would result.

[6] Usually, those erroneously tagged forms were imperfects mistakenly judged to be thematic aorists.

| LEMMA | AORIST STEM | AORIST TYPE | 1.sg.act.ind. | 2.sg.act.ind. | 3.sg.act.ind. | etc. |
|---|---|---|---|---|---|---|
| ἄγω | ἀξα- | A07 | 5 | 2 | 31 | … |
| ἀλέξω | ἀλεξα- | A07 | 2 | – | 2 | … |
| βαίνω | βη- | A01 | 7 | – | 180 | … |

Table 5.2: Example Data Entry for Homeric Aorists

I initially prepared the spreadsheet as alphabetized by lemma. In the original spreadsheet, I included all lemmata and recorded the token frequency of verbs with prepositional prefixes (preverbs) as well.[7] After the preliminary input of all data, I copied the data to a new spreadsheet, wherein I deleted all the lines of lemmata for prepositional verbs, if a corresponding simplex verb existed. I added the sum token frequency of that prepositional verb to the token frequency of the simplex verb. I undertook this procedure both in order to eliminate a large number of spurious hapax legomena in the results, and to reflect the fact that the derived aorist stem is, in a real sense, contained within prefixed forms.[8] For instance, the stem ἐξ-άψα- [eks-apsa-] 'attach', which is a hapax legomenon, shows the same sigmatic aorist stem as occurs in simplex ἅψα- [hapsa-] 'attach', which occurs $18\times$, and is therefore unlikely to be independent of the simplex; as its semantics would further indicate, the derived form with prefix is very close to its base. A compound verb with low token frequency probably even contributes to the exemplar representation of its simplex, which is why I add the tokens of the complex to the simplex. Likewise, the simplex βάλλω [bal:ɔ:] 'throw', has corresponding compound verbs ἀμφιβάλλω 'toss around', ἐκβάλλω 'throw at', etc., all of which have an A04 aorist βαλο- [balo-]; since the aorist stem proper in ἀμφιβαλο-, ἐκβαλο-, etc., is no different than simplex βαλο-, to count ἀμφιβαλο-, ἐκβαλο-, etc., as occurrences of an independent aorist stem would be misleading.

Moreover, all of the derivational prefixes found with Greek verbs appear to be highly productive, and therefore should be highly parsable and thus separable from their bases; the prefix ἀμφί- [amp$^{h}$í-] in Homer, for example, has $91$ hapax legomena and a $\mathscr{P}$ of $0.203$ (compare the $\mathscr{P}$ for the aorist formations in Table 5.3 below). Thus, even in cases in which a prefixed form exhibits a token frequency greater than its base,[9] if one follows the reasoning discussed under 4.3, to leave any prefixed verbal forms as independent types in the data would result in an overstatement of the number of truly distinct and independent types.

I also eliminated all iterative aorist forms with suffix -σκ- [-sk-] that I initially recorded – such forms are closely tied to, or even dependent upon, imperfects in [-sk-], and hence their position among aorists generally is dubious; see Chantraine 1958: 323–5 for further details. I removed a very small number of forms having passive aorist morphology (suffix -η-, -θη-, or

[7] In cases of tmesis (i.e., where a preverb surfaces at a distances from the verb itself, rather than as a prefix), it appears that the verb was tagged purely as belonging to the simplex lemma.

[8] Note, moreover, that very few cases in either Homer or the NT exist in which a prefixed verbal lemma exhibits an aorist formation different from the simplex; in those few cases, the prefixed verb with a distinct aorist stem was left as an independent entry. Furthermore, in cases in which a simplex is not present in the corpus, but multiple prefixed forms occur, the prefixed forms all always agree in the type of aorist.

[9] Only very rarely in Homer is the token frequency of the prefixed form appreciably greater than its base.

-θε/o-) that the Chicago Homer tagged as active or middle in function, since they fall outside the scope of the present investigation.

With this modified dataset, I could then sort the data by frequency and by aorist type, in order to extract the necessary statistical data: *V*, *N*, and *n*, for each respective (sub)type. At this phase, I also cross-checked the classification of some formations with other sources, and recategorized some forms that I had labeled erroneously during the original data entry.

## 5.3 Raw Results and Statistics

The revised dataset left some 774 distinct aorist stems, out of an originally recorded 1444, once I removed many stems following the procedure described above. The basic word distributions for the dataset are given in Table 5.3.

| *V* | *N* | Mean *N* | Median *N* | Mode *N* | Max *N* | Std.Dev. |
|---|---|---|---|---|---|---|
| 774 | 15500 | 20.02584 | 5 | 1 | 825 | 60.26407 |

Table 5.3: Basic Type and Token Statistics for Homeric Aorists

These statistics are typical of word distributions in large samples more generally: low median, large difference between minimum and maximum, mode of one, and large standard deviation. These values indicate that, as a subcorpus of Homer, this aorist dataset conforms to the statistical properties of natural language corpora generally, as discussed in Chapter 2, and is large enough on its own to render reliable results.

Table 5.4 below shows the distribution of types by subclass.

| **AORIST TYPE** | *V* |
|---|---|
| ATHEMATIC | **40** |
| A01 | 31 |
| A02 | 3 |
| A03 | 6 |
| THEMATIC | **117** |
| A04 | 69 |
| A05 | 20 |
| A06 | 28 |
| SIGMATIC | **615** |
| A07 | 526 |
| A07b | 75 |
| A08 | 7 |
| A09 | 7 |

Table 5.4: Distribution of Aorist Type Frequencies in Homer

| AORIST TYPE | $N$ | $n_1 \sim \mathscr{P}*$ | $\mathscr{P}$ | $n_1/V$ | $S$ | $\mathscr{I}$ |
|---|---|---|---|---|---|---|
| ATHEMATIC | 2735 | 1 | .0003656 | .025 | 40.58454 | 1.014613 |
| A01 | 1470 | 1 | 0.0006802 | .03226 | – | – |
| A02 | 1023 | 0 | 0 | 0 | – | – |
| A03 | 242 | 0 | 0 | 0 | – | – |
| THEMATIC | 6018 | 13 | .0021602 | .1 | 126.789 | 1.076923 |
| A04 | 3483 | 7 | .0020098 | .10145 | – | – |
| A05 | 1326 | 2 | .0015083 | .1 | – | – |
| A06 | 1209 | 3 | .00248 | .1071429 | – | – |
| SIGMATIC | 6749 | 158 | .0234108 | .2569106 | 802 | 1.304065 |
| A07 | 6022 | 138 | .02291597 | .2623574 | – | – |
| A07b | 653 | 17 | .02603369 | .2266667 | – | – |
| A08 | 18 | 3 | .16 | .4285714 | – | – |
| A09 | 56 | 0 | 0 | 0 | – | – |

Table 5.5: Productivity Statistics for Aorists in Homer

The overwhelming dominance of sigmatic aorist types is striking but fits the observation that sigmatic aorists are commonplace (Risch 1974: 246; $\sim$ 80% of all types), while root and thematic aorists are more restricted. The measures of productivity $\mathscr{P}$ and $\mathscr{P}*$ also fit the characterization of the sigmatic aorist as relatively more productive than root or thematic aorists. Recall that $n_1$ on its own can stand in for the exact calculation of $\mathscr{P}*$ (cf. Baayen 1993: 193), since the denominator (the total number of hapax legomena in the corpus) is constant across that corpus, so only the numerator (the type hapax legomena) is needed for comparison of $\mathscr{P}*$ internal to a single corpus. I also include the ratio of hapax legomena to types here, not because this measure has any demonstrable mathematical or psychological significance, but merely to show the proportion of types for which novel formations may be responsible. These statistics are given in Table 5.5.

The picture that these statistical measures provide is clear and unmistakable: sigmatic aorist formations are more productive in both the strict sense (the probability that the next sigmatic aorist sampled will be a hapax legomenon, i.e., $\mathscr{P}$) and in the global sense (the number of hapax legomena that the type contributes to the total number of hapax legomena in the sample, i.e., $\mathscr{P}* = n_1$); the ratio of hapax legomena to types is also high. We can straightforwardly conclude that sigmatic aorists are both highly parsable (given the comparatively large $\mathscr{P}$) and very useful; according to Powell (1988: 64), the text of Homer contains $1578$ hapax legomena, meaning that sigmatic aorists alone account for just over 10% of all unique stems in the text. More generally, these results are good in that they fulfill one of the requirements that Baayen and Lieber (1991: 809) give for a measurement of morphological productivity: "it reflects the linguist's intuitions concerning productivity." The outcome here is happy, because it matches the intuition of Greek scholars and Indo-Europeanists, that root aorists are decidedly recessive, whereas sigmatic aorists are productive. These productivity measures further make predictions about the historical development of the aorist as a cat-

egory, namely, that sigmatic aorists should come to make up an even greater proportion of aorist formations.

Plotting the vocabulary growth curves (VGCs) for each of the three major categories in Figure 5.1 neatly illustrates the same relationships between categories as in Table 5.5.[10] In the figure, two lines for each category are given, in which the upper line is the ratio of types to tokens, $V/N$, and the lower line is the ratio of hapax legomena to tokens, $n_1/N$; the latter thus plots the change in the estimate of $\mathscr{P}$ as more tokens are sampled. Note that the number of hapax legomena belonging to sigmatic aorists consistently exceeds the total number of types belonging to either thematic or root aorists.

Look now to the estimate of population types, $S$, for each category, and the corresponding measures of pragmatic potentiality, $\mathscr{I}$. In complete accord with the calculated values for productivity *stricto sensu*, $\mathscr{P}$, the estimates for $S$ obtained with a finite Zipf-Mandelbrot model barely exceed the sample types $V$ in the case of root and thematic aorists – the population of Greek root and thematic aorists appears to be practically exhausted in just the text of Homer. Sigmatic aorists, conversely, appear capable of generating a substantial number of new types: $S$ substantially exceeds $V$, and $\mathscr{I}$ is healthily above $1$. To make these points explicit, let us compare extrapolated vocabulary growth curves, for sigmatic and thematic aorists, just up to a sample of $20000$ tokens for each category (NB: the sample tokens $N$ are $6013$ and $6749$, respectively); this is Figure 5.2.[11] At these sample sizes, the crucial point to note is that the VGC for thematic aorists is already asymptotic (i.e., $\Delta V/\Delta N = 0$) beyond approximately $10000$ tokens, whereas the the VGC for sigmatic aorists is still rising ($\Delta V/\Delta N > 0$).

As I already emphasized in parts of Chapter 1, my objective in using these quantitative measures is not to refute existing claims about the productivity of certain morphological categories, but rather to make those claims more precise and meaningful by introducing scalar measures that permit the direct comparison of one formation to another. These first-order results precisely achieve that goal. In addition, this dataset lays the groundwork for a more detailed reckoning of both the synchronic and diachronic factors that have operated to produce the frequency distributions of these categories as they appear in Homer.

To obtain a better picture of how productive these categories are, one can make a rough comparison to results on the productivity of some nominal, adjectival, and verbal derivational affixes of English, as described in Baayen and Lieber 1991: 820 ff.. Because the numbers and statistics are not comparable in a strict sense, due to differences in the size of the corpora ($18000000$ words in the CELEX English corpus versus $\sim 200000$ words in Homer), this comparison is not wholly reliable. In general, however, given the behaviors of word distributions, we can note that $\mathscr{P}$ values for productive morphological processes will necessarily be smaller in larger corpora, because the larger size makes it more likely that the same word

[10] These VGCs are not empirical VGCs, obtained by sampling tokens from the linear sequence in which they occur in the text, but growth curves constructed from the frequency spectrum of each category (cf. Tables 2.3 or 2.4 for an example) through binomial interpolation.

[11] Given an approximately constant rate of growth for the number of tokens sampled in each category, obtaining a token sample of $20000$ for thematic aorists would require a corpus in the Homeric genre of $600000$–$700000$ words (the actual corpus contains $\sim 199000$ tokens; see 4.2.2 above).

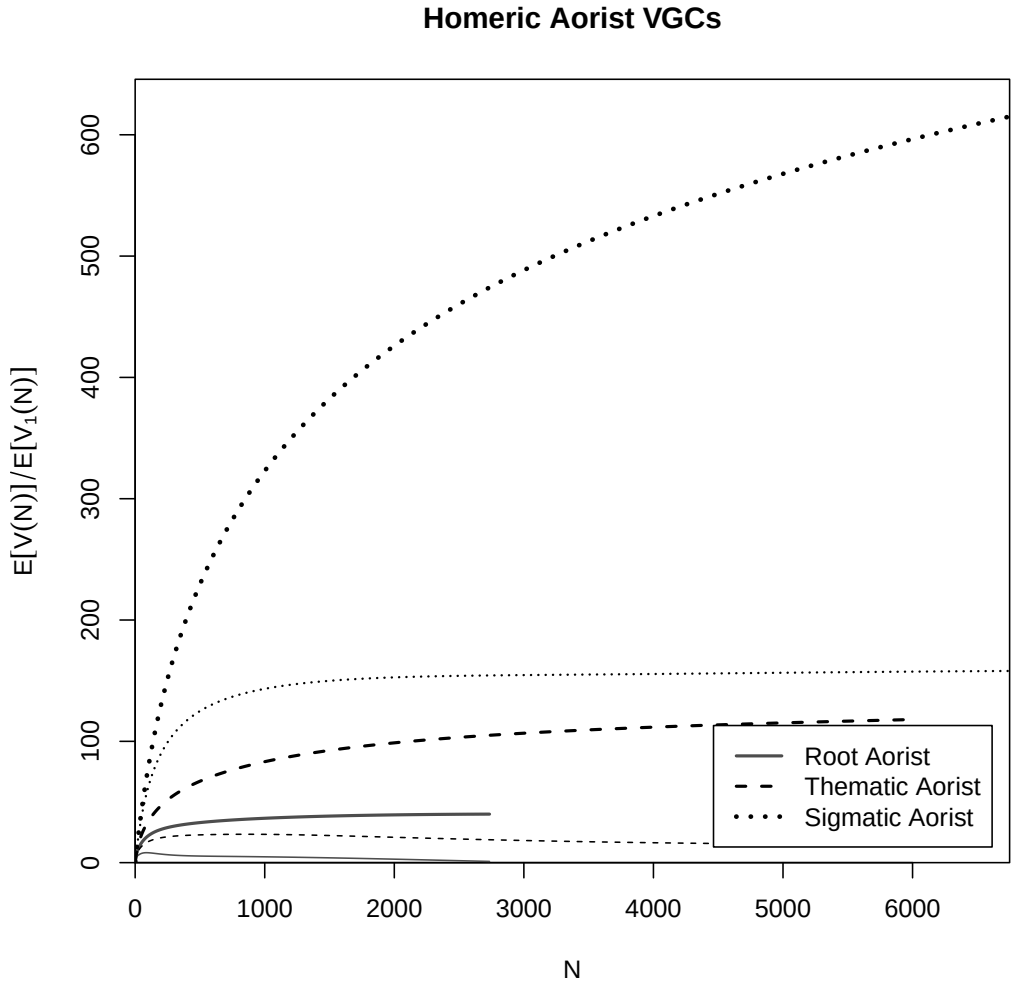



Figure 5.1: Vocabulary Growth Curves of Aorist Types in Homer

is sampled more than a single time. For example, the suffixation of English *-ness*, which represents a gold standard for a productive formation, has a $\mathscr{P}$ value of $.0044$ in the CELEX corpus. Meanwhile, unproductive processes, such as simplex nouns or nouns in *-al*, have a $\mathscr{P}$ value of less than $.0001$ in that corpus. If we accept that root aorists are unproductive, as they certainly are relative to thematic and sigmatic aorists, then we can deduce that categories with $\mathscr{P}$ values of $.001$ or less are to be considered wholly unproductive in the Homeric corpus. Moreover, the root aorists, as morphologically simplex stems, should approximate a lower bound for estimate of $\mathscr{P}$ for an category. Similarly, morphologically simplex nouns in Homer (non-deverbative "root nouns"; cf. Risch 1974: 3–5), which ought likewise to represent an inherently unproductive category, show a $\mathscr{P}$ of $.002021$ ($N = 3463$, $n_1 = 7$); this is less than the estimate of $\mathscr{P}$ for thematic aorists, and so allows us to refine the wholly

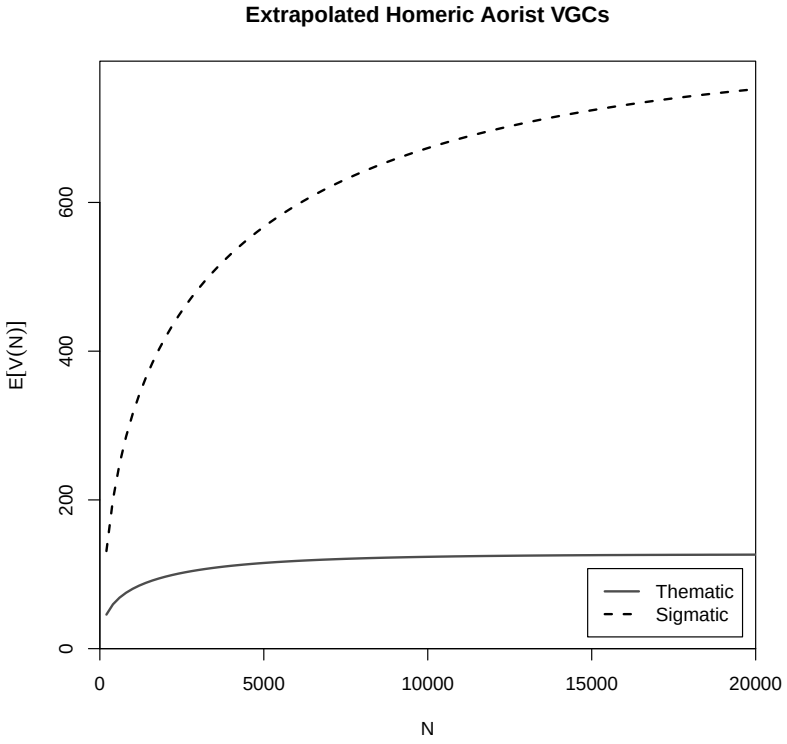



Figure 5.2: Extrapolated VGCs of Thematic and Sigmatic Aorists based on Homeric Data

unproductive range as somewhere between .001 and .002.[12]

In examining the subtypes of each major type, the results ($\mathscr{P}$ values) closely resemble the results for the larger category. Some subcategories of the sigmatic type, however, present some surprises. First is the apparent non-productivity of type A09, which shows a mixed sigmatic and thematic suffix. Risch (1974: 250) expresses the general view on this type: "Offenbar sind sie größenteils aus imperativisch verwendeten Futura entstanden."[13] Since this formation is limited to just a few types, and furthermore does not occur in Greek outside of epic poetry, its non-productivity here perhaps offers support for the notion that it is a peculiarity of Greek epic, and was not even readily used there. Conversely, type A08, with suffix -ησα- [-ɛ:sa-] seems to have an incredible degree of productivity. Among the aorist types that he sets up, this type is the only one that van de Laar (2000: xv) explicitly labels as "productive", and he later (pg. 411, fn. 15) declares that all A08 forms collected in his study "are secondary" (i.e., cannot possibly be forms with an inherited Indo-European basis). Nevertheless, the very small number of tokens found for this sub-type (18) means that the rate of

[12] It may be the case that root nouns, in comparison to root verbal formations, genuinely are more productive, given the fact that non-derived nouns may more readily enter the language as borrowings than non-derived verbs.

[13] "They have clearly arisen, in large part, from futures employed imperatively."

hapax legomena found in Homer has a higher chance of being influenced by sampling error. Yet perhaps the forms of this type in Homer may reflect the early, very productive days of a new type, before some of its particular forms have become established in the language and grown in token frequency.[14]

As a final piece of evidence for the stark contrast in productivity between sigmatic and root or thematic aorists, I have compiled the fifty types with highest token frequency in Table 5.6. In this group, sigmatic aorists constitute a minority $(19/50)$, which is striking in light in of how many sigmatic aorist types $(615/774)$ there are in the dataset. This sort of distribution anecdotally supports Baayen's (1989: 4) claim regarding "the importance of high token frequencies for types covered by unproductive word-formation rules in order to survive the pressure of rival, productive rules." The frequency distributions surrounding Homeric aorists appear entirely concordant with the hypothesis that forms instantiating unproductive morphological processes persist because their relatively higher token frequencies permit successful storage in and retrieval from the lexicon.

[14] Compare explicitly the situation of this subtype in the New Testament, where the A08 category appears to be moribund, attesting just a single, relatively frequent type, θελησα- [tʰelɛːsa-] (to (ἐ)θέλω 'will, want', 26×).

## 5.4 Analysis

With a clear picture of the degree to which each of the aorist types is productive, the primary synchronic problem is to establish what language internal factors may affect the degree of productivity. In effect, I will attempt to uncover some systemic motivations behind the degree of productivity. Knowledge of those factors ought to be indicative of how speakers derive these types of formations, which is in turn relevant for the history of those formations. For instance, if the relevant conditions for the productivity of a type are known not to have existed at an earlier period, then to reconstruct that type as being productive at an earlier phase is less well motivated. In particular, I think that the most compelling questions are the following two:

1. What synchronic patterns are present that make the sigmatic aorist the aorist formant *par excellence*, and why did those patterns emerge diachronically?

2. Given the, at best, marginal productivity of thematic aorist types, what situations obtained diachronically to produce the considerable number of thematic aorist types attested in Homer?

As a means of answering the first question, I will try here to establish the sort of morphophonological rules that could allow for the derivation of the attested aorist types, following the model of Albright and Hayes' work as described in 2.3.2. Recall that rules with high confidence are rules that can attract new members that meet their structural description, and likewise resist pressures of attraction from other rules. As a first step towards this goal, I will profile the hapax legomena of the three major types in 5.4.1, since precisely those forms are most likely to be generated by rule, rather than retrieved from memory; their morphological and phonological features may offer insight into what factors induce an aorist to be formed in one way or another. The second part of the first question, the origin of the sigmatic aorist's productivity, is fundamentally a question of why sigmatic aorists became, in effect, the default aorist stem for virtually all types of derived verbal stems. A look into the second question, meanwhile, I will delay until later in Chapter 6, once I have evaluated the cognate thematic aorist in Vedic, so that I can cast some comparative light on the category's diachronic developments.

### 5.4.1 Profiles of Formational Types

Profiles of the rare types (hapax and dis legomena) in each categories may help to give a sense of the domains (phonological, morphological, semantic) that allow for the coinage of new types belonging to each respective aorist formation. For the root and thematic aorists, the small number of hapax and dis legomena allows me to be fairly comprehensive, whereas the large number of sigmatic aorist hapax legomena requires more of a bird's-eye view.

| **LEMMA** | Gloss | **AORIST STEM** | **AORIST TYPE** | *N* | Absolute Rank |
|---|---|---|---|---|---|
| | | ATHEMATIC | | | |
| τίθημι | 'put' | θηκ- | A02 | 431 | 4 |
| βαίνω | 'come, go' | βη- | A01 | 428 | 5 |
| δίδωμι | 'give' | δωκ- | A02 | 360 | 8 |
| ἵστημι | 'stand' | στη- | A01 | 336 | 10 |
| ἵημι | 'release, throw, send' | ἡκ- | A02 | 232 | 12 |
| δύω | 'go into, go down' | δῡ- | A01 | 111 | 23 |
| γιγνώσκω | 'know' | γνω- | A01 | 92 | 28 |
| χέω | 'pour' | χευα- | A03 | 72 | 38 |
| κτείνω | 'kill' | κτα- | A01 | 61 | 45 |
| | | THEMATIC | | | |
| εἶπον | 'speak' | εἰπο- | A06 | 825 | 1 |
| ἔρχομαι | 'come, go' | ἐλ(υ)θο- | A04 | 723 | 2 |
| ὁράω | 'see' | ἰδο- | A04 | 479 | 3 |
| αἱρέω | 'take, grasp' | ἑλο- | A04 | 383 | 6 |
| βάλλω | 'throw' | βαλο- | A04 | 377 | 7 |
| ἱκνέομαι | 'reach, arrive' | ἱκο- | A04 | 352 | 9 |
| γίγνομαι | '(be)come [into being]' | γενο- | A04 | 248 | 11 |
| πίπτω | 'fall down' | πεσο- | A05 | 192 | 13 |
| ἔχω | 'have, hold' | σχο- | A04 | 167 | 14 |
| λείπω | 'leave' | λιπο- | A04 | 133 | 17 |
| λαμβάνω | 'seize' | λαβο- | A04 | 132 | 18 |
| ὄλλυμι | 'destroy' | ὀλο- | A05 | 125 | 21 |
| φεύγω | 'flee' | φυγο- | A04 | 108 | 24 |
| τίκτω | 'generate' | τεκο- | A05 | 106 | 25 |
| ἄγω | 'drive' | ἀγαγο- | A06 | 105 | 26 |
| εὑρίσκω | 'find (out)' | εὑρο- | A05 | 86 | 31 |
| θνήσκω | 'die' | θανο- | A04 | 75 | 33 |
| πάσχω | 'undergo, suffer' | παθο- | A04 | 64 | 42 |
| πόρον | 'furnish, offer' | πορο- | A05 | 61 | 46 |
| θείνω | 'slay' | πεφνο- | A06 | 60 | 48 |
| | | SIGMATIC | | | |
| φωνέω | 'speak (loudly)' | φωνησα- | A07 | 164 | 15 |
| ἐλαύνω | 'push' | ἐλασα- | A07 | 138 | 16 |
| νοέω | 'perceive, think' | νοησα- | A07 | 131 | 19 |
| ἀκούω | 'hear' | ἀκουσα- | A07 | 129 | 20 |
| ἐρύω | 'drag, pull' | ἐρυσσα- | A07 | 115 | 22 |
| λύω | 'release, free' | λυσα- | A07 | 103 | 27 |
| καλύπτω | 'cover' | καλυψα- | A07 | 89 | 29 |
| αἴσσω | 'shoot' | αἴξα- | A07 | 87 | 30 |
| ὄρνυμι | 'rise' | ὀρσα- | A07 | 82 | 32 |
| παύω | 'stop' | παυσα- | A07 | 74 | 34 |
| τελέω | 'accomplish' | τελεσσα- | A07 | 72 | 37 |
| ὄλλυμι | 'destroy' | ὀλεσα- | A07 | 70 | 39 |
| δέω | 'bind' | δησα- | A07 | 64 | 40 |
| καλέω | 'call' | καλεσσα- | A07 | 63 | 42 |
| δείδω | 'fear' | δεισα- | A07 | 62 | 43 |
| κελεύω | 'order' | κελευσα- | A07 | 61 | 44 |
| ὠθέω | 'push' | ὠσα- | A07 | 61 | 47 |
| ποιέω | 'make' | ποιησα- | A07 | 60 | 49 |
| λέγω | 'pick up; say' | λεξα- | A07 | 59 | 50 |

Table 5.6: 50 Highest-Frequency Aorist Stems in Homer

#### 5.4.1.1 Root Aorists

Table 5.7 gives the complete data on the hapax, dis, and tris legomena of the root aorists, all of which belong to the Ao1 type:

| LEMMA | STEM | GLOSS | *N* | FORMS | Citation |
|---|---|---|---|---|---|
| πέρθω | περ- | 'devastate' | 1 | inf.mid. πέρθαι | *Il.* 16.708 |
| ἀναπάλλω | (ἀνα)παλ- | 'spring, swing' | 2 | 3.sg.mid.ind. ἀνέπαλτο | *Il.* 8.85, 23.694 |
| ἀποδιδράσκω | ἀποδρα- | 'flee' | 2 | part.act.nom.sg.m. ἀποδράς | *Od.* 16.65, 17.516 |
| λέγω | λεγ- | 'pick up; say' | 2 | 3.sg.mid.ind, 1.sg.mid.ind. λέκτο, ἐλέγμην | *Od.* 4.451, 9.335 |
| ῥύομαι | ῥυ- | 'protect, guard' | 2 | 3.pl.mid.ind. ῥύατ(ο) | *Il.* 18.515, *Od.* 17.201 |
| σβέννυμι | σβη- | 'extinguish' | 2 | 3.sg.act.ind. ἔσβη | *Il.* 9.471, *Od.* 3.182 |
| ἀραρίσκω | ἀρ- | 'fit together' | 3 | part.mid.acc.sg.m. ἄρμενον | *Il.* 18.600, *Od.* 5.234, 5.254 |
| πλέω | ἀποπλω-; ἐπιπλω- | 'sail' | 3 | 3.sg.act.ind.; 2.sg.act.ind. ἀπέπλω; ἐπιπλώς, ἐπέπλως | *Il.* 16.708; *Il.* 6.291, *Od.* 3.15 |

Table 5.7: Athematic Aorist Hapax, Dis, and Tris Legomena in Homer

This set of infrequent root aorists exhibits a phonological profile similar to other root aorists: they belong to roots, from the IE point of view, in final sonorants (παλ- [pal-] < **pl̥-*, δρα- [dra-] < **dr̥-*) or laryngeals (πλω- [plɔː-] < **ploh$_{1/3}$-* (cf. OE *flōwan* 'flow', σβη- [sbɛː-] < *[zg$^{w}$esh$_{2}$-]). By my reckoning, only four of the 40 root aorist types are formed to roots ending in final non-sonorant consonants; eight of the 40 root aorist stems (30 Ao1 stems) belong to resonant final-roots; 17 (and all three of the Ao2 type, i.e., ablauting δωκ- [dɔːk-], ἡκ- [hɛːk-], and θηκ- [t$^{h}$ɛːk-]) belong to laryngeal-final roots, and especially the tendency for IE laryngeal-final (Greek long-vowel) roots to have root aorists is notable.[15]

The crucial and obvious fact that does distinguish root aorists in general from the other two aorist categories is that they are not, from either a synchronic or etymological perspective, derived formations. The verbal root, without any other potentially analyzable derivational material, builds the aorist stem. From the perspective of the verbal system as a whole, root aorists normally correspond to present stems with (historically) analyzable derivational material, such as reduplication or other suffixes; following van de Laar (2000: 379–80), the apparent Greek examples of root presents alongside root aorists are largely, perhaps entirely, spurious. Given that this type of aorist is an underived formation, for it to be productive would be difficult: either new lexical roots would have to somehow enter the language, or speakers would have to abandon the formal differentiation of present and aorist stems.

[15] We will see in Chapter 6 that the pairing of long-*ā* roots with root aorists appears to hold in Vedic Sanskrit as well.

If indeed root aorists should have no degree of productivity whatsoever, is the fact that even a couple of hapax legomena occur surprising? In part, the $n_1$ here may be an artifact of the relatively small corpus size. Nevertheless, already in this corpus, the small $\mathscr{P}$ value is good evidence that new root aorists would be very rare creatures. On the other hand, such low token frequency forms here might reflect archaisms, in case their tokens have largely been replaced by the corresponding productive category. Except for the stems ἀποδρα- [apodra-] 'fled', ῥυ- [ru-] 'protect, guard', and πλω- [plɔ:-] 'sail', a competing sigmatic aorist built to the same root is already present in Homer for all of the verbs in Table 5.7.

In fact, the single hapax legomenon, the middle infinitive πέρθαι [pérth ai], in fact, is not beyond doubt. Chantraine (1958: 384) mentions the suggestion a suggestion of Meillet that the form rests on a haplologized present infinitive *πέρθεσθαι [pérth est^h ai], functioning as a present infinitive. This suggestion is self-evidently unprovable, though the very fact that πέρθαι [pérth ai] represents the sole instance of such a root aorist in Greek, while corresponding thematic (πραθο- [prath o-], $9\times$ Hom.) and sigmatic (περσα- [persa-], $29\times$ Hom.) aorists seem well established, and the very unproductivity of root aorists as a whole offers a reason to doubt the form's reality. Harðarson (1993: 207) judges that the form is a sigmatic aorist, /pert^h-s-st^h ai/ → [pérth ai], though as the Attic form πέρσαι ← /pert^h-s-ai/ indicates, the phonological derivation of /pert^h-s-st^h ai/ → [pérth ai] will not work without introducing new assumptions surrounding the development of clusters involving dentals and [s].

Given the number of present stems that do exhibit potentially analyzable derivational material, the fact that root aorists are so few and unproductive is perhaps surprising, since a morphophonological rule of the form (X)-$\sqrt{}$-(X) → (Ø)-$\sqrt{}$-(Ø) / #_#, i.e., simply deleting the derivational affixes of the present stem within the boundaries of the word, could easily generate root aorist stems. The only necessary proviso would be that the speaker be equipped to analyze the root from the derivational affixes of the present stems in all cases. This operation is arguably well-instantiated only for reduplicated presents of the type δίδωμι [dídɔ:mi] 'give', where the ratio of number of such presents with root aorists versus number of such presents with sigmatic aorists in van de Laar (2000: 326–7) is $7:6$. That is, despite the productivity of sigmatic aorists, presents with reduplication still often correspond to root aorists. The correspondence between reduplicated present formations and root aorists, however, rests mainly on types with very high token frequency: five of the eight reduplicated presents corresponding to root aorists occurring in Homer (γιγνώσκω [gignɔ́:skɔ:], ἵστημι [hístɛ:mi], δίδωμι [didɔ:mi], τίθημι [tít^hɛ:mi], and ἵημι[híɛ:mi]) occur among the fifty most frequent aorist types (see the table above), and two others (πίμπλημι [pímplɛ:mi] and ὀνίνημι [onínɛ:mi]) have frequency well above the median (14 and 7 vs. 5). Yet (ἀπο)διδράσκω [(apo)didráskɔ:] occurs only twice.[16] Note that, among the high frequency types here, parallel sigmatic aorists do not occur in Homer (the sigmatic aorist στησα- [stɛ:sa-] is a causative 'make stand', in contrast to the intransitive root aorist στη- [stɛ:-] 'stand'), whereas those of the lower frequency items do have parallel sigmatic aorists in Homer. Thus, in agreement

[16] The only occurrences of the verb ἀποδιδράσκω [apodidráskɔ:] in Homer are limited to the two root aorist forms, and the base διδράσκω [didraskɔ:] does not occur in the corpus at all. One might therefore conclude that, despite being a low-frequency verb, the aorist stem ἀποδρα- [apodra-] is not productively derived because it is the most frequent stem within its verbal lemma.

with the view of Hay and Baayen 2003, it seems that the high frequency types are immune to parsing (or the construction of a morphophonological rule for their derivation), but at the same time resist pressure of productive, well-instantiated processes; lower frequency types may not be so fortunate.

#### 5.4.1.2 Thematic Aorists

Table 5.8 gives the complete data on the hapax legomena among the thematic aorists:

| LEMMA | STEM | GLOSS | *N* | FORMS | Citation |
|---|---|---|---|---|---|
| ἅλλομαι | ἁλο- | 'jump' | 1 | 3.sg.mid.subj. ἅληται | *Il.* 21.536 |
| ἀνακράζω | ἀνακραγο- | 'cry out' | 1 | 1.sg.act.ind ἀνέκραγον | *Od.* 14.467 |
| ἐπιπίπτω | ἐπιπτο- | 'fall upon, attack' | 1 | inf.mid. ἐπιπτέσθαι | *Il.* 4.126 |
| ἐρείκω | ἐρικο- | 'rip' | 1 | 3.sg.act.ind. ἤρικε | *Il.* 17.295 |
| κεύθω | κυθο- | 'cover, hide' | 1 | 3.sg.act.ind κύθε | *Od.* 3.16 |
| κρίζω | κρικο- | 'creak, screech' | 1 | 3.sg.act.ind. κρίκε | *Il.* 16.470 |
| κυδαίνω | κυδανο- | 'glorify' | 1 | 3.pl.act.ind. ἐκύδανον | *Il.* 20.42 |
| γοάω | γοο- | 'groan' | 1 | 3.pl.act.ind. γόον | *Il.* 6.500 |
| κεράννῡμι | κερο- | 'mix' | 1 | 3.pl.subj.mid. κέρωνται | *Il.* 4.260 |
| κεύθω | κεκυθο- | 'hide, cover' | 1 | 3.pl.act.subj. κεκύθωσι | *Od.* 6.303 |
| λαμβάνω | λελαβο- | 'seize' | 1 | inf.mid. λελαβέσθαι | *Od.* 4.388 |

Table 5.8: Thematic Aorist Hapax Legomena in Homer

Among the varieties of thematic aorist, from the phonological point of view, only class A04 forms a relatively coherent group: almost all of the roots involved in the class contain a liquid, nasal, or high vowel, thus making the root amenable to taking zero-grade ablaut (from an Indo-European perspective). Among this group of hapax legomena, however, no features mark them as a phonologically coherent group. Four verbs concern some noisemaking ('cry out', 'creak', 'groan', 'call upon'), though this fact is probably a coincidence.

More interesting, however, is the fact that several of these thematic aorist stems attest other aorist stems as well in Homer:

- beside Ao4 ἁλο- [(h)alo-], Ao1 ἁλ- [(h)al-] (57×).
- beside Ao4 (ἐπι)πτο- [(epi)pto-], Ao5 πεσο- [peso-] (192×).
- beside Ao4 κυθο- [kut$^{h}$o-], Ao6 κεκυθο- [kekut$^{h}$o-], Ao7 (ἐπι)κευσα- [(epi)keusa-] (1×).
- beside Ao6 λελαβο- [lelabo-], Ao4 λαβο- [labo-] (132×).
- beside Ao4 κερο- [kero-], Ao7 κερασσα- [kerassa-] (9×).

In all of these cases, we find a unique formation, sometimes peculiar to Homer, alongside the aorist formation that regularly appears in the Greek of the 1st millennium BCE. The case of Ao4 κυθο- [kut$^{h}$o-] and Ao6 κεκυθο- [kekut$^{h}$o-] is particularly unusual, as both are unique in Greek to their respective attestations in the *Odyssey*. The usual aorist to κεύθω [keut$^{h}$ɔ:] in Classical Greek is sigmatic, though κευσα- [keusa-] happens to occur in Homer only in the prefixed form ἐπικευσα- [epikeusa-] (1×).

More interesting is the co-occurrence of an Ao4 alongside an Ao6, as in the case of λαβο- [labo-] and λελαβο- [lelabo-] or κυθο- [kut$^{h}$o-] and κεκυθο- [kekut$^{h}$o-]. The relevant data appears in Table 5.9. This pattern recurs with several other verbs, which have evident morphological and phonological similarities: on the one hand, several form double nasal presents (like λαμβάνω [lambanɔ:), for which the presence of an Ao4 is a reliable pattern, and otherwise show a root-final θ [t$^{h}$]. Only ἐνίπτω, with Ao4 ἐνῖπτο- [eni:pto-] and Ao6 ἐνῖπαπο- [eni:papo-], stands outside of either one of those patterns. That verb more generally is very rare outside of Homer and the later imitators of epic poetry; though the Ao4 appears to have a genuine appearance in Strabo (*Geographica* 13.1.7.62), the Ao6 is to be found nowhere else.

| **LEMMA** | **GLOSS** | Ao4 | Ao6 | Ao4 *N* | Ao4 *N* |
|---|---|---|---|---|---|
| ἐνίπτω | ‘reprove, upbraid’ | ἐνῖπο- | ἐνῖπαπο- | 15 | 6 |
| κεύθω | ‘hide, cover’ | κυθο- | κεκυθο- | 1 | 1 |
| λαγχάνω | ‘obtain’ | λαχο- | λελαχο- | 21 | 4 |
| λαμβάνω | ‘seize’ | λαβο- | λελαβο- | 132 | 1 |
| λανθάνω | ‘escape notice, miss’ | λαθο- | λελαθο- | 41 | 9 |
| πείθω | ‘persuade’ | πιθο- | πεπιθο- | 53 | 13 |
| πυνθάνομαι | ‘become aware ’ | πυθο- | πεπυθο- | 43 | 3 |

Table 5.9: Instances of Ao4 alongside Ao6 in Homer

Indeed, then, for all of these verbs, the reduplicated aorist looks like a feature of Homer and epic poetry. The appropriate question, then, is whether these reduplicated aorists, for which κεκυθο- [kekut$^{h}$o-] and λελαβο- [lelabo-] provide the only instances of hapax legomena in the category, reflect the genuine, but old, members of a moribund category, or are rather inventions of the epic *Kunstsprache*.[17] If the latter response is the right one, then perhaps these types should be dismissed, and then the reduplicated aorist would appear to be

[17] See Chantraine 1958: 395–8 on the contrasting function of the Ao6 opposite the Ao4.

a yet more restricted and moribund category, on the level of root aorists. Nussbaum (1987), for instance, suggests that reduplicated aorists forms like λέλαχον [lélak$^{h}$on] are built on the stem of the reduplicated perfect, which is attested for all of these verbs in Homer, with the exception of ἐνίπτω [enípt͡ɔː], again clearly an aberrant member of this set, and λαμβάνω [lambán͡ɔː]. This conception is also upheld in the detailed study of Beckwith (1996) into the Greek reduplicated aorist, and thus, following Beckwith, I am inclined to be suspicious of the linguistic reality for most of the A06 forms listed above.

#### 5.4.1.3 Sigmatic Aorists

The number of hapax legomena among the sigmatic aorist types is too numerous to display conveniently. However, examining unique sigmatic aorist stems in this corpus casts light on the productivity of all the aorist categories. In considering just the hapax legomena to class A07, one finds that some 80 of the 158 are built to verbal stems that are themselves derived (viz., denominal or deadjectival verbs). These derived verbal stems form 1.sg.pres.act.ind. forms in -ίζω, -έω, -άω, and -όω; some are denominatives derived with the suffix *-i̯e/o- (e.g., ἀνάσσω [anass͡ɔː] 'rule' < *u̯anak-i̯e/o-, cf. (ϝ)ἄναξ [(w)anaks] 'ruler'), others are causatives with *o* grade of the root (e.g., δοκέω [doke͡ɔː] 'seem'), and yet others deverbative (e.g., δεικανάω [deikanáː] 'welcome, greet'), to name some derivational roles that these groups fulfill. In contrast, root and thematic aorists corresponding to derived presents in -ίζω, -έω, -άω, and -όω simply do not exist.

In effect, much of the measurable productivity of the sigmatic aorist is attributable to the fact that all of these derived verbal stems obligatorily form their aorists using the suffix -σ(α)-. While granting that not every present in -ίζω, -έω, -άω, and -όω is non-primary (cf., e.g., τορέω [toréː] 'pierce', which cannot be derived in Greek), a large majority are, and 235 of the 615 sigmatic aorist types look to correspond to such presents. If all such forms were removed from the dataset, sigmatic aorists would then show 111 hapax legomena and 435 types, giving a $n_1/V$ ratio of .25; compare the $n_1/V$ ratio of .345 for sigmatic aorists above. Hence, the proportion of sigmatic aorist hapax legomena to derived verbal stems is disproportionate to the number of sigmatic aorists to derived verbal stems in the dataset; a $\chi^2$-test indicates that this distribution is statistically significant at the 5% level.[18] Thus, while sigmatic aorists would still appear to be the most productive aorist type even if no secondarily derived verbal stems occurred in the data, the presence of those derived stems helps to explain their overwhelming productivity. See Tucker (1990) for a systematic treatment of all the Homeric vowel-stem verbs; the author systematically discusses the likely inner-Greek or Indo-European sources for all such stems.

The genuine morphological interest of the sigmatic aorist, beyond its great productivity, concerns the reliability and extensibility of mappings to the aorist stem from other parts of the verbal lemma, especially the present stem. These patterns can be most perspicuously

[18] $\chi^2 = 3.89, p = .048$. Note that the distribution is probably even more skewed, and so more strongly significant, than the data as reported here suggest, because some of the 235 sigmatic aorists corresponding to presents in -ίζω, -έω, -άω, and -όω certainly are non-derived, while I have confirmed that 80 of the sigmatic aorist hapax legomena are indeed built to derived verbal stems.

identified and discussed within the Minimal Generalization Learning framework, to which I now turn.

### 5.4.2 Minimal Generalization Learning of Present : Aorist Patterns

Although corpus-based measures used above provide credible data concerning the productivity of the various aorist types, they do not clearly illustrate the word-formation rules that come into play when constructing those aorist stems. While for types with high-token frequency storage in memory likely obviates the need for online generation, the types with low token frequency do require some sort of online production. For the production and processing of those low frequency types, a rule is necessary; the problem is to discover the rules at work. One should bear in mind that differences in token frequency indeed neatly distinguish the major varieties of aorist:

| Mean $N$ | Median $N$ |
|---|---|
| Root | |
| 68.375 | 24.5 |
| Thematic | |
| 51.342 | 9 |
| Sigmatic | |
| 10.953 | 4 |

The usual description of verbal stems in Greek and Vedic in Indo-Europeanist practice is to set up the stems as derivatives to a root. In many cases in Vedic, and even more so in Greek, no pure root formation is synchronically available in the language. Hence, one must assume either that speakers carry out a great deal of abstraction in the construction of lexical entries (e.g., pres. δρέπω [drepɔ:] 'pluck' and aor. δρέψα [drepsa] fall under an entry for a root /drep-/, from which the respective stems are derived by suffixes), or that some actually occurring stem may serve as a base of derivation for other stems.

In case studies concerning paradigmatic analogies in Latin, Lakhota, and Yiddish, Albright (2002b) proceeds from surface bases (i.e., actually occurring word forms): single paradigm members act as the base for the other paradigm members. The parallel situation at the level of derivational morphology (as opposed to the inflectional level, which Albright has studied), would be for one stem to serve as a base of derivation for other stems.

Given these assumptions, I undertook a test of how learnable the pattern mappings between the aorist stems attested in Homer and their corresponding present stems might be, using the Minimal Generalization Learner as described in Chapter 2. Specifically, I treated the present stem as the base, from which the aorist stem would then be derived. Morphophonological mappings with high confidence that the MGL is able to discover should be indicative of potentially productive sub-patterns.

A general comparison of present and aorist stems suggests that the present would generally serve as a better base just in case the aorist is sigmatic, because the [s] suffix induces a large number of phonological neutralizations: all root-final labials are reduced to [p] (e.g.,

pres. λείβω [leíbɔː] : aor. λειψα- [leipsa-], pres. τρέφω [trépʰɔː] : aor. θρεψα- [tʰrepsa-]), all root-final dentals to Ø (e.g., pres. ψεύδω [pseúdɔː] : aor. ψευσα- [pseusa-]), and all root-final velars to [k] (e.g., pres. ὀρέγω [oregɔː] : aor. ὀρεξα- [oreksa-]). Conversely, because thematic aorists induce no phonological neutralizations to the root, and only some slots of the root aorist paradigm do so, those aorist stems are less neutralizing than their corresponding present stems; yet, a survey of the corresponding presents to root and thematic aorists indicates that only rarely is the present stem built with a formant (namely, the suffix */-i̯e/o-/) that would be phonologically neutralizing. Hence, it is specifically the pairing of a simple thematic present stem (derived with the suffix */-e/o-/) and a sigmatic aorist where the mapping present → aorist is most beneficial.

One complicating issue in modeling is the fact that the learning data derived from Homer often presents multiple aorist stems in correspondence with the same present stem. In principle, the MGL is able to learn multiple different rules that could potentially apply to an input datum (in this case, a present stem), and assess their likelihood of application in the form of rule confidence. Instances of present stems that correspond to multiple aorist stems are thus preserved as such in the learning data, assuming that learners indeed faced variant outputs in some cases; the multiple options may serve to support different distinct rules within the system.

#### 5.4.2.1 Data Preparation

In preparing the dataset for a learning run, I started from the pared list of 774 aorist stems; I then eliminated from the list all present : aorist pairs which were suppletive (e.g., pres. ὁράω [oraɔː] 'see' : aor. ἴδον [idon], pres. φέρω [pʰerɔː] 'bring' : aor. ἐνείκα [eneːka]), since such patterns would be unlearnable in any case, and only add possible confounds by weakening the reliability of numerous rules. The prepared file was left with 749 present stem → aorist stem learning pairs as training data; the same present stems were given as testing data.[19] These files and information on their encoding are available at:
`https://github.com/rpsandell/SandellDiss`.

In order to compensate for the fact that the MGL cannot discover discontinuous changes, I represented all reduplication as a suffix `R` immediately following the verbal root, thus making the reduplication continuous with any suffixes. Present stems belonging to the thematic conjugation were represented with `o-`: for example, the dictionary lemma γιγνώσκω is represented as `gnvRsko-` in the learner file. The input mapping of pres. `gnvRsko-` to root aorist `gnv-` would then, in principle, be learnable as a change `Rsko` → Ø. Meanwhile, athematic presents were represented with simply a dash following the stem: for example, the dictionary lemma ὀνίνημι is represented as `onHR-`, or the lemma ὄλλυμι as `ollu-`, in the learner file; root aorists were represented with simply a `-` following the stem; thematic aorists were represented with `o-`; sigmatic aorists were represented with `sa-`, `a-`, `Hsa-`, and `so-` (A07, A07b, A08, and A09, respectively).

[19] Ultimately, the output returned a prediction for only 746 forms; three mappings in the learning data represented unique non-recurrent mappings over which no rule with a scope of at least 2 could apply.

I also prepared a file containing the phonological features for each segment used in the input file, excepting the R used to represent reduplication. The use of phonological features permits the formulation of rules in the form of more general featural relationships, thereby capturing patterns that go beyond purely segmental identity.

#### 5.4.2.2 Results and Evaluation

The results, in terms of the sheer number of different morphophonological rules discovered, demonstrate the heterogeneity of present stem → aorist stem mappings. The MGL ultimately preserved $6804$ distinct rules, of which only $220$ have a scope of $100$ or greater. This behavior is typical of the MGL when presented with large and diverse datasets; the MGL similarly preserves thousands of rules when fed the 2181 English present : past tense pairs that composed the basis of the study in Albright and Hayes 2003. The rules learned for this Greek data show that the correct generation of most forms indeed rests on many diverse Islands of Reliability, with a scope of between 3 and 30 forms: among "winning" rule applications, the mean scope is $23.583$, and more than half of the "winning" outputs $(535/746)$ rely on rules with a scope of $20$ or less. Indeed the rules with the worst reliability and confidence are precisely those with very large scopes. Among the rules with a scope greater than $100$, only rules showing the change [eo-] → [ɛːsa-] (i.e., presents like ἀγνοέω [agnoeɔː] : aor. ἀγνοήσα [agnoɛːsa]) and [o-] → [sa-] / $XY_{[+high]}\_$ have a reliability greater than $0.5$.

I evaluate the performance of the grammar constructed by the learner along two parameters: learnability and accuracy. Learnability in this context refers to the possibility of inducing some rule from the training data that, when applied to the same input, will generate the same output as found a given input → output mapping. For instance, if the grammar contains a rule that can generate the output [agnoɛːsa-] when given the input [agnoeo-], then the mapping [agnoeo-] → [agnoɛːsa-] is learnable. Accuracy is the extent to which the grammar correctly reproduces the training data, i.e.., when the rule with the best confidence that takes scope over an input generates the same output as in the training data. For example, the rule with the best confidence that applies to the input [agnoeo-] indeed predicts [agnoɛːsa-], and thus the mapping [agnoeo-] → [agnoɛːsa-] is correctly learnable.

As a procedure, the performance of the learner's ability to reproduce the data is evaluated in the following manner:

- the learner applies every rule that could take scope over a test form, and generates the output.
- the possible outputs from a test form are sorted according to the confidence score of the rule.
- the form generated by the highest-confidence rule is considered the "winner", i.e., the form that this micro-grammar should generate.
- if the "winning" form is the same as the form in the learning data, then the form is correct.

| Total Tests | **Learnable** | **Unlearnable** | **Correct** | **Incorrect** |
|---|---|---|---|---|
| | | Root | | |
| 36 | 23 | 13 | 4 | 32 |
| | | Thematic | | |
| 108 | 72 | 36 | 28 | 80 |
| | | Sigmatic | | |
| 607 | 607 | 0 | 507 | 100 |

Table 5.10: MGL Performance on Homeric Present : Aorist Stem Mappings

- If the "winning" form differs from the form in the training data, then the form is incorrect.
- if the correct form (the form that appeared in the training data) does not appear as a possible output from any rule applied to the test form, the form is considered unlearnable.

Results were obtained through the electronic comparison of lists of the forms to be predicted and lists of possible outputs and winning outputs generated by the learner. Winners were obtained by sorting the MGL output by confidence and input number, then selecting the first instance of each input number in the table. The results of the evaluation appear in the Table 5.10.[20]

In part, the dataset itself is biased towards the learning of sigmatic aorist forms, because they are so much more numerous in terms of types – the mappings between present stems and sigmatic aorist stems have potentially hundreds of supporting instances, whereas patterns mapping to thematic or root aorists are simply represented by many fewer types. However, because very general rules perform so poorly in this dataset, some effective Islands did emerge even among non-sigmatic aorists. For instance, the MGL appropriately discovered the kappatic athematic aorists (A02, ἵημι : ἧκα, τίθημι : θῆκα, δίδωμι : δῶκα; [híɛːmi] : [hêːka], [títʰɛːmi] : [tʰêːka], [dídɔːmi] : [dɔ̂ːka]) and established the rule Red $\rightarrow$ [k] / $V_{[\text{-ATR}]\_}$ as an Island applying to the presents of those forms. The correspondence between a reduplicated present and plain root aorist (A01, ἵστημι : στῆ, πίμπλημι : πλῆ; [hístɛːmi] : [stêː], [pímplɛːmi] : [plɛː]) is also recognized in a rule Red $\rightarrow$ Ø, though its confidence fails to make it the "winning" rule for such forms. "Double nasal" presents (e.g., μανθάνω [mantʰánɔː], λανθάνω [lantʰánɔː], πυνθάνω [puntʰánɔː]) corresponding to A04-type thematic aorists made up another Island that was able to resist the application of more general rules that would produce sigmatic aorists. Indeed, dictionary lemmata show that such double nasal presents normally exhibit only thematic aorists, and not sigmatic aorists, in Ancient Greek. It thus seems that

[20] In preliminary phases of research, before the list of aorist forms was thoroughly vetted and pared, a simulation was run without the use of phonological features learned, which learned many fewer rules, just 293. In performance, the simulation showed a comparable degree of accuracy (67% correct, vs. 71% with features). The use of features seems most to improve performance among the thematic aorists, where accuracy increases from less than 10% without the use of features to over 25% with features.

the "double nasal" → thematic aorist Island does successfully project even some token frequency thematic aorists during the 1st Millennium BCE.

In general, winning forms are projected with a reasonable degree of confidence: the mean confidence of the winner is $0.6798601$. Among sigmatic patterns alone, this rises to $0.7261852$; among thematic aorists, this falls to $0.4853577$, and among root aorists, is similar at $0.4820193$. Overall, however, the performance of the MGL, as I anticipated, is abysmal for root and thematic aorists: not only does the learner predict incorrect outputs, but the correct output was, in about $45\%$ of cases, unlearnable. Most of the incorrect outputs among sigmatic aorists result from one of three types of errors:

- extension of the rules [eo-] → [ɛːsa-] / X_ and [ao-] → [ɛːsa-] / X_ to forms that show -εσα- [-esa-] or -ασα- [-asa-] rather than -ησα- [-ɛːsa-] in their aorists.
- extension of a rule [dzo-] → [sa-] / X_, where the application of [dzo-] → [ksa-] / X_ would have generated a correct output.[21]
- the application of rules that generate Ao7b-type forms (e.g., [eira-] → [eːsa-] / X_) to inputs that instead have an Ao7 (e.g., predicting [keːra-] as the best outcome over [kersa-] from pres. [keírɔː]).

All three of these learning errors find some historical support. The extension of the patterns [eo-] → [ɛːsa-] / X_ and [ao-] → [ɛːsa-] / X_ over [eo-] → [esa-] / X_ and [ao-] → [asa-] / X_ is evident in the history of the language. For example, the aor.inf.act. κένσαι [kénsai] 'to goad' (*Il.* 23.337) is perhaps the last gasp of a primary sigmatic aorist to the root **kent-*; the stem κεντησα- [kentɛːsa-] (pres. κεντέω [kentéɔː]) normally appears after Homer (e.g., Sophocles *Ajax* 1245). On the other hand, some primary sigmatic aorists to roots in -έω [-éɔː], such as ζέσσεν [dzésːen] (*Il.* 18.349, *Od.* 10.360) to ζέω [dzéɔː] 'cook, boil' (PIE */i̯es-/, cf. Skt. *yasati* 'boils, heats'), never ultimately fall into the vowel-lengthening pattern in the aorist (but beyond relatively early works, the aorist to ζέω occurs almost exclusively in the active participle, though a 3.sg.act.ind. ἔζεσεν [édzesen] occurs in the Septuagint). Similarly, among presents in -ζω [-dzɔː] in the New Testament (see 5.5 below on the New Testament Data), nearly all show an aorist in -σα- [-sa-], rather than -ξα- [-ksa-]. Likewise, outside of Homer, the sequence of sonorant + [s] in aorist forms is almost non-existent – only the Ao7b type is licensed to roots terminating in sonorants.

These results together unswervingly confirm the productivity of the sigmatic aorist, and would point towards extension at the expense of the other two types. Indeed, many thematic or root aorists show sigmatic aorists as well already in Homer (e.g., Ao7b κτεινα- [kteːna-] alongside thematic Ao4 κτανο- [ktano-] and root Ao1 κτα- [kta-], or Ao7 (ἐπι)πλωσα- [(epi)plɔːsa-] alongside Ao1 (ἀπο)πλω- [apoplɔː-]); about half of all Ao1 have a sigmatic aorist (not including semantically motivated cases like βη- 'went' versus βησα- 'made go'). In the

[21] That is, the learner tended to see the mapping of present in -ζω [-dzɔː] to aorist in -σα- [-sa-] as a more reliable and extensible pattern than present in -ζω [-dzɔː] to aorist in -ξα [-ksa]. Alternations in the form of a sigmatic aorist found in Homer, such as ἤρπαξε/ἥρπασε [hɛːrpakse/hɛːrpase] 'seized' reflect exactly the ambiguity in mapping and possibility of both [-sa-] and [-ksa-] outputs from the same input.

historical course of the Greek language, perhaps a more interesting question than why sigmatic aorists replace other types, is why some aorists should resist "sigmaticization" at all.

### 5.4.3 Other Evidence of Productivity

Two other correlates of productivity mentioned in Hay & Baayen's research (as discussed in 3.4 above) have not received attention in this case study: the ratio between the derived form and its base, and the role of phonotactic frequency in helping to establish morpheme boundaries. For the purposes of this particular study, base : derived ratios would be difficult to assess, since, although I have treated the present stem as a base in the MGL study above, what exactly one should consider to be the synchronic base of stems that appear to be primary formations is difficult to establish. This problem could, unfortunately, be recurrent in the study of many Greek and Vedic nominal and verbal formations, because so many specific forms, which are morphologically complex from at least an etymological perspective, are synchronically without bases. If, however, such "baseless" formations tend to correlate with non-productivity, then they may provide evidence that their bases did indeed exist historically – the bases must have been available in order to permit the formation of those (now unproductive) derivatives in the first place. Nevertheless, as further examination of the Greek aorist in section 5.5 will show, some fairly clear evidence emerges that the present stem should act as the base within a verbal lemma in Greek.

The role of phonotactics in motivating the greater parsability of sigmatic aorists, in comparison to thematic aorists, meanwhile, is largely self-evident. Namely, the consonant clusters that appear at the root-suffix boundary in many sigmatic aorists (e.g., ψ [ps], ξ [ks]) are less common than corresponding consonant + ε/ο [e/o] sequences that appear at the root-suffix boundary in thematic aorists. Furthermore, such sequences of consonants inevitably translate into a syllable boundary, whereas a CV sequence instead forms a syllable itself. In sigmatic aorists then, a prosodic boundary is typically well-aligned with the morphological boundary, whereas those boundaries do not align in the case of thematic aorists. All things considered, the common alignment between syllable and morpheme boundary ought to provide an edge in parsability for sigmatic aorists.

### 5.4.4 Conclusions concerning the Homeric Aorists

As stated at the outset of this chapter, the general conclusions about aorist formations in Homer are unsurprising, in that they agree with intuitive observations that follow from reading the texts. Sigmatic aorists, according to both the statistics $\mathscr{P}$, $\mathscr{P}*$, and $\mathscr{I}$ are very productive: such aorists are more parsable, more useful, and have greater pragmatic potentiality, dominating the other types along each of these dimensions. The MGL simulation further established that most of the particular sigmatic aorist patterns found in Homer are readily learnable on the basis of their corresponding present stems. Conversely, the statistics $\mathscr{P}$, $\mathscr{P}*$, $\mathscr{I}$ indicate that root aorists are effectively moribund (as would be expected, in principle, for any morphologically simplex category), and thematic aorists are at best marginally productive; the MGL simulation demonstrated that most patterns for building root and the-

matic aorists are less reliable than sigmatic aorist patterns, or are altogether unlearnable. The co-occurrence of sigmatic aorists alongside root or thematic aorists in Homer is thus an expected consequence of the fact that productive formations will oust corresponding unproductive forms that high token frequency does not protect. The fact that intuition, statistical measures, and learning simulations all converge on the same result should inspire confidence that corpus-based statistics and learning simulations can help to determine a category's degree of productivity in cases in which intuition is a less reliable guide. Statistics and learning simulations also serve to quantify intuition in cases in which intuition captures the general pattern. One can then discuss changes in productivity as changes in grammar proper, because statements of the form "category X has become more/less productive by degree Y" become possible.

The question at this point is: how do these facts concerning productivity practically benefit research into the history of Greek and Indo-European? As I already discussed in section 1.3.2, and above in 3.4, very low $\mathscr{P}$ value below a certain threshold is essentially a guarantee that the forms pertaining to that category cannot be synchronically generated. The question then shifts to the matter of how old, exactly, the form must be. On this point, the consideration of individual lexical frequency may be helpful. In the list of the highest token frequency aorists above, all of the root aorists given there, except for δῡ- 'went into', seem securely reconstructible on the basis of cognate forms. Among the thematic aorists, the stems having the highest token frequency, ἐλ(υ)θο- [el(u)t$^{h}$o-] 'arrived, started' and εἰπο- [eipo-] 'spoke', similarly find confirmation in cognate forms (Ved. *ruhá-* and OIr. *-luid* [lud$^{j}$], Toch. B *lac* [ləts]; Ved. *vóca-*, Av. *vaoca-*), from apparent Indo-European formations */h$_{1}$lud$^{h}$-o-/ and */u̯e-u̯k$^{w}$-o-/; I will revisit this point at the close of Chapter 6.[22] In contrast, none of the highest frequency sigmatic aorists know of matching cognate formations. The relevant reasoning seems to be as follows:

1. productively formed (young) lexical items have low token frequency (by definition);
2. lexical items having high token frequency are resistant to replacement by productively built forms;
3. therefore, lexical items having high token frequency are likely to be older than lexical items having low token frequency.

In practice, this reasoning is complicated by the fact that lexical replacement will be a gradual, rather than instantaneous process. For example, Homer still shows the root aorist (ἐπι/ἀπο) πλω- [(epi/apo)plɔ:-] 'sailed' (*Il.* and *Od.*), as well as sigmatic aorist πλωσα- [plɔ:sa-] (*Il.* only); in the period of the composition of the Homeric epics, πλωσα- evidently has not wholly replaced πλω-. The fact, however, that πλω- [plɔ:-] was ripe for replacement, whereas δωκ-

[22] On both the formal comparative basis and the basis of comparative frequency (Ved. *ruhá-* 31× RV, well above median frequency of RVic aorists), to claim that the thematic aorists to the root */h$_{1}$leu̯d$^{h}$-/ are "post-Indo-European thematizations or innovations", as does Rix et al. (2001: 20; 248), is not credible. This judgment felicitously accords with the conclusions of Cardona (1960: 123), that */h$_{1}$leu̯d$^{h}$-e/o-/ was one of at least two thematic aorists that existed in PIE.

[dɔːk-] ‘gave’ was not, may be partially a function of the lower token frequency that πλω- [plɔː-] had. This effect is comprehensible from the point of view of language processing, as treated in 3.2 and 3.3: because πλω- [plɔː-] was a lexical item of low token frequency, to access πλω- [plɔː-] from memory would have required more time than generating the sigmatic aorist πλωσα- [plɔːsa-] online; conversely, the high token frequency of δωκ- [dɔːk-] allowed for rapid lexical access, thus rendering the production possibility of a competing *δωσα- [dɔːsa-] very small. The fact that δωκ- falls into a (weak) morphological Island of Reliability may have offered some further protection against creeping sigmatism as well.

At the level of individual lexemes, I can then express two generalizations:

1. the less productive that a morphological category is, the more likely that one of its members is old.

2. the higher the token frequency of a lexical item, the more likely that it in particular is old.

In this fashion, the interpretation of low token frequency items for both productive and unproductive categories would be consistent: among productive categories, rare items reflect recent creations, or even neologisms, while among unproductive categories, low token frequency items might comprise more recent additions to that category, which were more weakly established in use before the category began to fall into disuse.

How to apply these principles with respect to demonstrably productive categories presents an additional challenge: sorting what is necessarily new from what is possibly old. The most frequent among the sigmatic aorists reflect a number of formations that are plainly inner-Greek creations: aorists such as φωνησα- [pʰɔːnɛːsa-] ‘spoke’[23] and ἀκουσα- [akowsa-] ‘heard’ both lack comparative support for a sigmatic aorist specifically. Meanwhile, the stem ῥεξα- [reksa-] (ῥέζω [redzɔː] ‘do, perform’), a primary sigmatic aorist, is less frequent than either of the just-mentioned aorists, though still common in absolute terms ($57\times$ Hom.). While its token frequency and lack of any competing aorist formations suggest that it is not the youngest of sigmatic aorists, at the same time, it is not likely to be the original Indo-European aorist either, given the OAv. evidence for a root aorist (2.sg. *varəš* ‘you did’ $<$ */u̯érĝ-s/). What is desirable to know then, is when, in the history of Greek, the sigmatic aorist was sufficiently productive to have supplanted the root aorist */u̯erĝ-/, so as to result in a sufficiently frequent sigmatic aorist that did not compete with a root aorist any longer in the Greek of the 1st Millennium BCE. The frustrating difficulty in dealing with a productive category is that it not only easily takes in new members, but it easily maintains existing members. Exactly how dependable a window lexical token frequency into history may be likewise remains a problem in need of better definition and quantification.

[23] See in particular Tucker 1990: 165; 190 on the historical derivation of this aorist stem.

## 5.5 The Greek Aorist after Homer: The Aorist in the Koiné

The long attested history of Greek offers the opportunity to directly examine and measure changes in the productivity of morphological processes. As an example of such work, I consider the situation of the same aorist categories examined in Homer, but in the predominant variety of Greek attested approximately eight centuries later, namely, the Koiné ("common, shared") Greek, as attested in the New Testament (NT), dating to the second half of the 1st century CE. In this section, I statistically evaluate the productivity of aorist types in the NT just as for Homer above, and then directly compare behaviors at the levels of both the categories as a whole and that of individual lexical items. With respect to the categories, there is little genuine change, as far as my methods can discern. Individual lexical items likewise exhibit remarkable stability, although some amount of change in derivational category in favor of sigmatic aorists is discernible; more notable is the simple disappearance of numerous root and thematic aorists, without any apparent renewal.

### 5.5.1 Data Collection: New Testament

As a corpus, the New Testament is appreciably smaller than Homer, containing just $137783$ tokens in comparison to the $\sim$ $199000$ of Homer. Thus, all other things being equal, the NT will provide less data about the categories under examination. Nevertheless, the overall quantity of available data (see the following subsection) is substantial enough that sampling errors do not appear to pose a grave concern.

My objective was to construct a database of aorist forms, tagged for subtype, just as was done with the Homeric material. The basis for this work was the recent electronic edition of the NT prepared by Holmes 2010, which includes embedded morphosyntactic tags for every token. This textual edition and its tagged data are easily accessible through software available from Logos Bible Software (`www.logos.com`).

Using that software interface, I was able to obtain a list of all (non-passive) aorist tokens in the NT, and output that list as an .xlsx file. I subsequently read all the .xlsx file into R as a dataframe. Using a series of small scripts in R, I then summed the number of identical forms, and output the resulting dataframe as a .txt file; each line thus identified the lemma, form, and frequency of that form. I subsequently collapsed together all forms sharing the same lemma and summed the frequencies of the forms; the resulting dataframe was likewise output as a .txt file. The following phase was to tag each line in that .txt file with the type of aorist stem attested, while also collapsing and summing the frequencies of forms with preverbs together with their respective simplexes.

### 5.5.2 Raw Results and Statistics: New Testament

The same procedures of data sorting and type reduction as applied to the data in Homer were applied to the NT forms. This procedure left some $595$ distinct aorist stems from an original set of 956. Basic observational statistics, as was the case for the Homeric aorists,

reveal that the subcorpus exhibits the properties of natural word frequency distributions, and so indicate a reliable subcorpus. In accordance with the simple fact that the NT corpus is smaller than the Homeric corpus, nearly all of these frequency statistics are also smaller; only the standard deviation, surprisingly, is slightly greater. These statistics appear in Table 5.11. These statistics include $257$ tokens of "alpha-thematic" aorists (i.e., thematic aorists that show a theme vowel [a] rather than [e/o]), which are excluded from further consideration because such forms arguably do not belong to Homer at all.[24,25]

| *V* | *N* | Mean *N* | Median *N* | Mode *N* | Max *N* | Std.Dev. |
|---|---|---|---|---|---|---|
| 595 | 9500 | 15.96639 | 3 | 1 | 810 | 62.07847 |

Table 5.11: Basic Type and Token Statistics for New Testament Aorists

The set of productivity statistics is given in Table 5.12. One small remark on the number of root aorist hapax legomena is necessary here. A 3.sg.act.ind. ἔδυ [edu:] may or may not occur at Mark 32.1; Holmes (2010) prints ἔδυ, but some older editions of the NT (e.g., Westcott and Hort 1881) print a sigmatic aorist ἔδυσεν [edu:sen]. The form occurs in the phrase ἔδυ(σεν) ὁ ἥλιος 'the sun set', which, with the root aorist, occurs identically as such $7\times$ in the Septuagint, and a similar variant ἔδυ φάος ἠελίοιο 'the light of the sun set' occurs in Homer. ἔδυ in the NT is certainly the *lectio difficilior*, against the sigmatic aorist that occurs there $26\times$, and given its embedding in an idiom, may well be the correct original form. The statistics as calculated below, however, use just a single hapax legomenon among the athematic aorists.

At first glance, some statistics would appear to tell a very different story from the statistics given for the Homeric data in Table 5.5. In particular, the calculation of $\mathscr{P}$ looks substantially larger for both root and sigmatic aorists in comparison to Homer. Bear in mind, however, that precisely because the NT corpus is smaller than the Homeric corpus, both fewer tokens will be sampled, and more tokens proportionally will appear as hapax legomena. Other procedures are needed to compare productivity statistics from corpora of different sizes.

### 5.5.3 Analysis

At 2.2.3.2, I argued that reducing the sample size $N$ to the $N$ of the smaller category as a means of comparing productivity (as advocated in Gaeta and Ricca 2006) amounts to throwing away useful data on frequency distributions that may have genuine psycholinguistic impact. In order to compare the productivity of the same category, where that data is drawn from two separate and independent corpora, however, the procedure of $N$-reduction is not

[24] In effect, "alpha-thematic" forms reflect the use of inflectional endings marking person and number as found on sigmatic aorists on stems that are otherwise identifiable as thematic aorists, e.g., 1.sg.act.ind. εἶπα [eipa] 'I spoke' rather than εἶπον [eipon]. How exactly to model the creation of such forms is an interesting question in itself.

[25] West (1998: XXX–I) does not print any "alpha-thematic" forms in his edition of the Iliad, though forms such as εἶπας [eîpas] for εἶπες [eîpes] are fairly profuse in manuscripts of Homer.

| AORIST TYPE | $V$ | $N$ | $n_1 \sim \mathscr{P}*$ | $\mathscr{P}$ | $n_1/V$ | $S$ | $\mathscr{I}$ |
|---|---|---|---|---|---|---|---|
| ATHEMATIC | 10 | 914 | 1 | .00109409 | .025 | 10.86154 | 1.086154 |
| A01 | 5 | 345 | 1 or 2 | .002898551 | .4 | – | – |
| A02 | 3 | 511 | 0 | 0 | 0 | – | – |
| A03 | 2 | 58 | 0 | 0 | 0 | – | – |
| THEMATIC | 34 | 3694 | 7 | .0024945 | .1271186 | 166.7208 | 4.903553 |
| A04 | 24 | 2089 | 6 | .002872188 | .25 | – | – |
| A05 | 8 | 697 | 1 | .00143472 | .125 | – | – |
| A06 | 2 | 908 | 0 | 0 | 0 | – | – |
| SIGMATIC | 544 | 4635 | 161 | .03473571 | .2959559 | 699.5728 | 1.285979 |
| A07 | 513 | 4129 | 151 | .0365706 | .294347 | – | – |
| A07b | 30 | 480 | 10 | .02083333 | .3 | – | – |
| A08 | 1 | 26 | 0 | 0 | 0 | – | – |

Table 5.12: Productivity Statistics for Aorists in the New Testament

only licit, but necessary, when the two corpora themselves are of different sizes. The correct fashion in which to evaluate the respective productivities of sigmatic aorists between Homer and the New Testament would be to draw sigmatic aorist tokens from Homer up to $N = 4635$, and then calculate $\mathscr{P}$. Since drawing a sample $4635$ sigmatic aorist tokens from Homer is not technically feasible, given the form of my data, the number of types and hapax legomena can instead be approximated at $N = 4635$ through binomial interpolation from the empirical frequency spectrum. Plotting the vocabulary growth curves resulting from interpolation of the sigmatic aorist data from Homer and the New Testament yields Figure 5.3.

Note in the figure that the difference in number of types is relatively small, never greater than about $20$ at any given point, and more importantly, the number of hapax legomena is also very much comparable. Binomial interpolation, for the Homeric data, estimates $n_1$ at $156.1497$, and thus $\mathscr{P} = 0.03368926$ (=$\frac{156.1497}{4635}$); $\mathscr{P}$ calculated for the sigmatic aorists in the NT was $.03473571$, about 3% greater. On the whole, we may conclude that the productivity of sigmatic aorists has remained relatively stable across eight centuries, perhaps slightly increasing.

The same procedure, when applied to the data on thematic and athematic aorists, makes clear that the former category has suffered a marked decrease in productivity from its already marginal degree of productivity in Homer. The relevant VGCs are given in Figures 5.4 and 5.5, where the substantially lower curves for $n_1$ in the NT in both figures suggest lesser productivity of the categories in the NT.

Even more remarkable is the precipitous decline in the number of types and tokens in the non-sigmatic categories in the NT; the comparison between Homer and the NT is summarized in Table 5.12. In terms of types, non-sigmatic types represented about $20\%$ of all types in Homer; this has fallen to under $8\%$ in the NT.

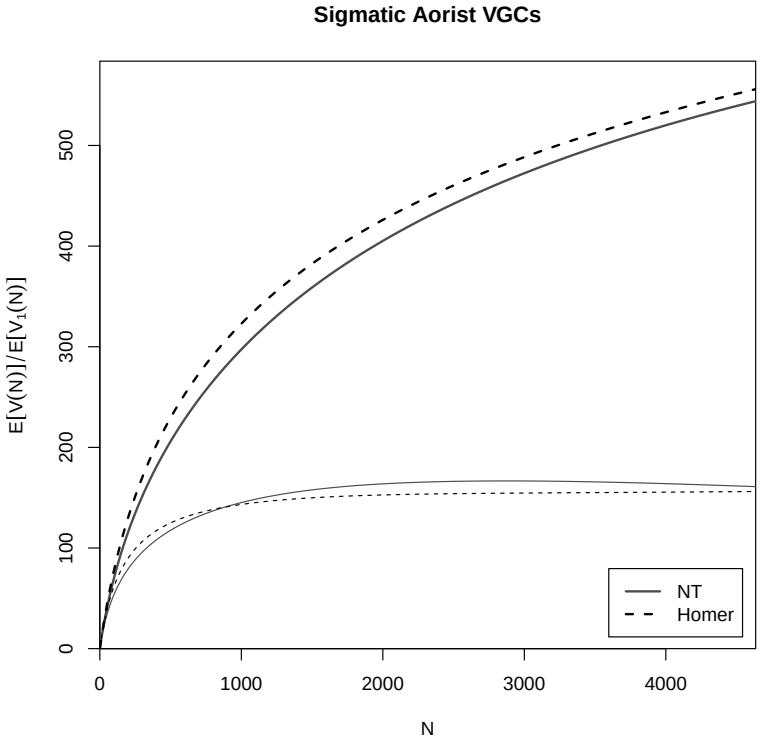



Figure 5.3: Interpolated Growth Curves of Sigmatic Aorists in Homer and the NT

| | Homer $V$ | Homer $N$ | NT $V$ | NT $N$ |
|---|---|---|---|---|
| Athematic | 40 | 2735 | 10 | 914 |
| Thematic | 117 | 6018 | 34 | 3694 |

Table 5.13: Comparison Type and Token Frequencies of Non-Sigmatic Aorists

Comparison of specific aorist stems between Homer and the NT shows that, in fact, there are relatively few clear instances of outright category changes (these instances appear in Table 5.13), but instead many athematic and thematic aorists present in Homer are simply absent from the NT. In Table 5.14, I draw the reader's attention in particular to the creation of the sigmatic aorist πιμπλησα- [pimplɛ:sa-] 'filled', which transparently illustrates derivation of the aorist stem from an originally reduplicated present stem.

The transition between Homer and the NT looks rather undramatic, in part because the same verbal root often attests multiple aorists already in Homer: e.g., κτείνω [kteínɔ:] 'kill' in Homer attests a root aorist κτα- [kta-], thematic aorist κτανο- [ktano-], and sigmatic aorist κτεινα- [kte:na-]. In virtually all such cases, it is the sigmatic aorist that remains in the NT: κτεινα- [kte:na-] appears there, but not κτα- [kta-] or κτανο- [ktano-]. Thus the decline in non-sigmatic aorist types and tokens evinced in the NT reflects the triumph of sigmatic op-

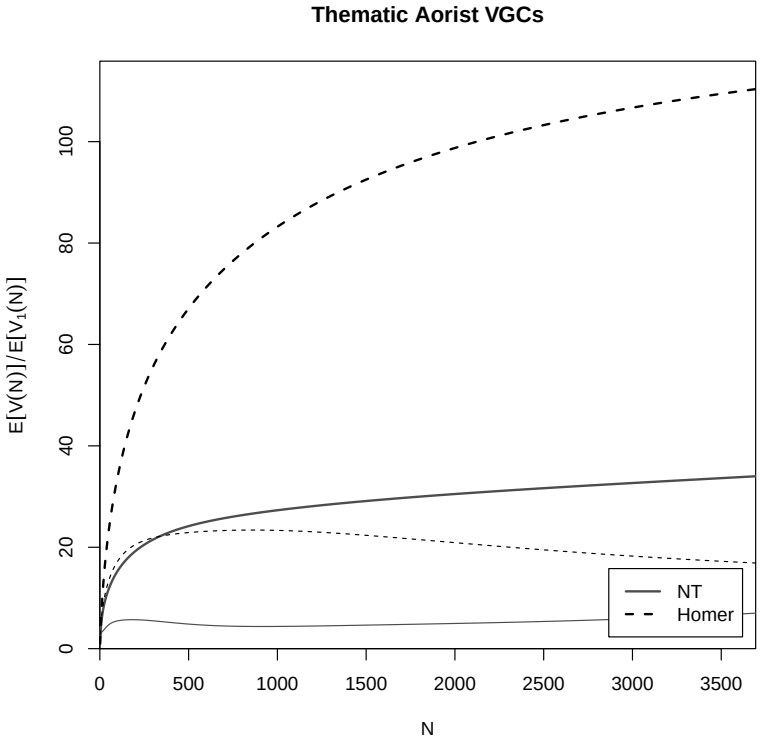



Figure 5.4: Interpolated Growth Curves of Thematic Aorists in Homer and the NT

| **Lemma** | Gloss | Homer Stem | Homer Type | NT Stem | NT Type |
|---|---|---|---|---|---|
| πίμπλημι | 'fill' | πλη- | A01 | πιμπλησα- | A07 |
| πλέω | 'sail' | -πλω- | A01 | πλευσα- | A07 |
| φθάνω | 'overtake' | φθα- | A01 | φθασα- | A07 |
| καίω | 'burn' | κηα- | A03 | -κευσα- | A07 |
| σεύω | 'hunt, drive away' | σευα- | A03 | -σευσα- | A07 |
| κράξω | 'shout, shriek' | κραγο- | A04 | κραξα- | A07 |

Table 5.14: Clear Category Changes between Homer and the NT

tions that were already available in Homer. Beyond the six instances given in Table 5.14 in which Homer attests only a non-sigmatic aorist that appears as sigmatic in the NT, I count 42 cases in which a sigmatic aorist exists alongside another non-sigmatic type in Homer, and the sigmatic version persists into the NT. Such sigmatic forms were also systematically the "winning" forms in the MGL simulation carried out above. In short, the snapshot of the Greek aorist system furnished by Homer shows that sigmatic aorists had already vanquished their competitors; the subsequent history would seem to consist in mopping up the survivors from the battlefield.

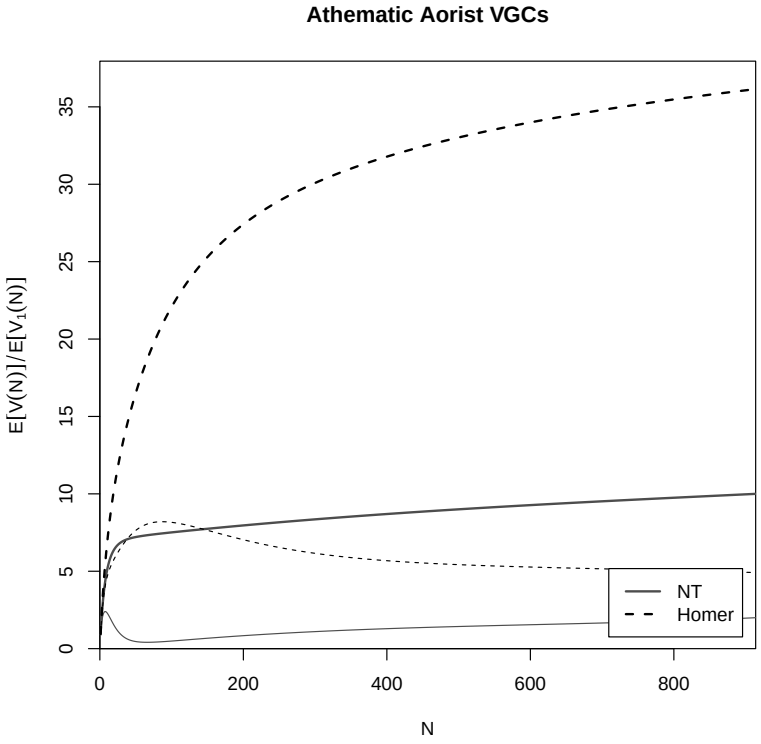



Figure 5.5: Interpolated Growth Curves of Athematic Aorists in Homer and the NT

## 5.6 Summary and Conclusions

The study undertaken in this chapter has demonstrated the successful application of the corpus-based productivity measures described in Chapter 3 to data drawn from relatively small corpora. Furthermore, I have shown that the consequences for unproductive categories, measured as such in an earlier corpus, in fact come to their logical conclusions in a later corpus. This work captures the two extremes of the diachronic spectrum; ideally, one should fill in some points in between, to trace the overall diachronic trajectory with greater precision. Consequently, to add the data on aorist formations from an author of the Classical period, e.g., Plato, would be desirable; if indeed a gradual decline in membership of the non-sigmatic categories is at work, then one would expect to find somewhat more types, and perhaps measure a higher degree of productivity among thematic aorists, at such an intermediate stage.

# CHAPTER 6

# The Aorist in the *Ṛgveda*

In the preceding chapter, I have demonstrated that the corpus-based measures of productivity introduced in Chapter 2 indeed successfully distinguish the principal varieties of Greek aorist formations in both Homer and the later New Testament. Most crucially, those measures unambiguously indicate that the sigmatic aorist was already a robustly productive category in the Greek of the early 1st millennium BCE, and comparison to the NT shows that little changed in this regard over the course of the centuries separating those two samples of the Greek language, beyond the expected attrition of non-sigmatic aorist types.

This chapter turns now to the productivity of the cognate categories in Vedic Sanskrit, based on the data furnished by the *Ṛgveda* (RV). In comparison to the Greek data, the Vedic data is more problematic in several ways: the corpus itself (probably) reflects a wider diachronic spectrum (and hence a series of more distinctly differentiated grammars than we find in Homer), and the corpus is slightly smaller (cf. 4.2 above). Furthermore, the content and style of the RV may result in making the aorist, in general, a less "useful" (in Baayen's sense) morphosyntactic category than in Homer, which further restricts the range of the data. Nevertheless, the quantifiable productivity of the distinct aorist formations is broadly similar to the situation found in Greek. An explicit comparison between Homeric Greek and RVic Sanskrit on this point would suggest that the last common ancestor of these two languages contained at least the seeds for a burgeoning productivity of sigmatic aorists.

## 6.1 Morphological Characterization

The description and classification of aorist aspectual stems in Vedic Sanskrit is, in broad outline, entirely parallel to the major aorist categories treated in the preceding chapter. Thus, once again, virtually all aorists may be described as ROOT/ATHEMATIC, THEMATIC, or SIGMATIC aorists.[1] Once again, just as in Greek, aorist stems are inflected for both active and middle voice, as well as a full range of moods, though participles are appreciably less common in Vedic. The formal description of the three types is also essentially the same: root aorists show no derivational suffix, thematic aorists show a suffix /-a-/, and sigmatic aorists a suffix containing an /s/. The only substantial formal difference between Vedic and Greek is that sigmatic aorists typically exhibit ablaut of the root in Vedic, with /ā/ in the active

[1] In Sanskrit grammatical terminology, it is usual to speak of the *s*-aorist rather than sigmatic aorist, but I will use here the term 'sigmatic aorist' in order to maintain terminological alignment with the Greek categories.

| | |
|---|---|
| Root Aorist | (∼ Greek A01–3) athematic, asigmatic, ablauting (e.g., *adām* 'I gave') |
| Thematic Aorist | (∼ Greek A04, A05) thematic, asigmatic, non-ablauting (e.g., *ávidam* 'I found') |
| Reduplicated Aorist | (∼ Greek A06) usually thematic, asigmatic, with CV reduplicant (e.g., *avocam* 'I spoke', *ápaptat* 'he flew') |
| Causative Aorist | usually thematic, asigmatic, with CV(ː) reduplicant (e.g., *ajījanat* 'he brought into being') |
| | Sigmatic |
| *s*-Aorist | (∼ Greek A07(b)) athematic suffix *-s*, ablauting (e.g., *ábhārṣam* 'I bore') |
| *iṣ*-Aorist | athematic suffix *iṣ*, sometimes ablauting (e.g., *akāniṣam* 'I enjoyed') |
| *siṣ*-Aorist | athematic suffix *siṣ*, non-ablauting (*ayāsiṣam* 'I implored') |
| *sa*-Aorist | thematic suffix *sa*, non-ablauting (e.g., *avṛkṣam* 'I tore') |

Table 6.1: Classification of Vedic Aorist Types

indicative/injunctive[2] (e.g., 1.sg.act.ind. *ahārṣam* 'I took') and /a/ or zero grade elsewhere (3.pl.mid.ind. *ahṛ̥ṣata* 'they took (for their own benefit)'), whereas the verbal root in Greek sigmatic aorists never ablauts. For further information on the details of aorist inflection, see Macdonell 1910: 365–85 or Macdonell 1916 [1993]: 158–75; other general resources on Indo-European verbal inflectional also rely substantially on the Vedic material, and may be profitably consulted. See also Narten 1964 (which is an important work of reference for the present undertaking) for detailed treatment and analysis of individual sigmatic aorists. For recent treatments of the syntax and semantics of the Vedic aorist see Dahl 2010: Ch. 4 (focused on the RV) and Dahl 2013 (on the functional development of the aorist through the history of Vedic).

While the general aorist categories are the same as in Greek, Vedic contains its own set of subtypes, some of which are not directly matched by any Greek formation. I give the classification scheme that I have adopted, with representative examples, in Table 6.1.

Vedic also possesses a causative aorist, which, because its characteristic morphology also involves reduplication, is often collapsed in grammatical description with the genuine non-causative reduplicated aorist. Formally, causative aorists require that the syllable corresponding to the reduplicant be bimoraic ("heavy"); in the case of a root containing two consonants at its left edge (e.g., $\sqrt{\textit{śrath}^{i}}$ 'pierce'), a causative and a non-causative aorist would be formally indistinguishable, and therefore forms such as *śiśrathat* must be evaluated based on their function and meaning in the text to establish their most appropriate classification. While the non-causative reduplicated aorist has an Indo-European background (with at least one impeccable cognate set: Ved. *voca-* : Av. *vaoca-* : Gk. εἶπον [eîpon] < /u̯e-u̯k$^{w}$-e/o-/), the causative aorist is a category innovated in Indic (cf. Leumann 1962, discussion in Jamison 1983: 214–9). The Vedic "passive aorist", despite its name, is a category both formally and

[2] 'Injunctive' is simply the traditional grammatical "mood" given to aorists used without the past tense marking prefix /á-/.

functionally largely separate from the aorist proper, and is not considered whatsoever in the following.

The most notable difference between the Greek and Vedic aorist formants is the proliferation of new sigmatic formants in Indic: the *iṣ*-, *siṣ*-, and *sa*-aorists are all novel creations.[3] The /-iṣ-/ is straightforwardly the result of a new morphemic analysis of what was, in origin, an allomorph of /s/: *iṣ* (normally) surfaces on *seṭ* roots[4] and simple *s* on *aniṭ* ('without *i*') roots. Already in the RV, however, /iṣ/ can be established as a distinct derivational morpheme, because it is used to build aorists to *aniṭ* roots: the original allomorphy of *iṣ* and *s* has broken down (and, indeed, simple *s* aorists are found to *seṭ* roots as well). *sa*-aorists, Narten (1964: 102) suggests, emerge from treating the subjunctive of a sigmatic aorist (i.e., /-s-a-/) as the stem of the indicative as well. The (type-wise) rare *siṣ*-aorist appears to be an *s*-aorist extended with the *iṣ*-aorist formant, e.g., 1.sg. /a-yā-s-iṣ-am/ → *ayāsiṣam* 'I went', though how and why such morphology should have emerged is somewhat mysterious. The emergence and development of these newer varieties of sigmatic aorist is a problem that could be examined both with respect to the quantitative productivity of those sub-categories and the morphological processes that could have given rise to them.[5]

## 6.2 Data Collection and Preparation

The same basic frequency information on all aorist forms occurring in the *Ṛgveda* (types, tokens, hapax legomena) as was necessary in the study of Greek aorists, is required here. To obtain data from the RV, my basic tool was the concordance of Lubotsky 1998; Lubotsky helpfully labels all stems that he judges to be aorists as such, so data could be gathered simply by reading through the concordance from cover to cover. Lubotsky's concordance has been occasionally supplemented and cross-checked with data from Grassmann 1872 [1976], and data on sigmatic aorists specifically was often checked against Narten 1964 (though Narten does not consistently provide detailed frequency data). Forms whose interpretation as an aorist that struck me as suspect were evaluated more thoroughly against their textual background, with the help of the commentary by Oldenberg 1909–12, and translations of Geldner 1951 and Jamison and Brereton 2014. Since, again as in the study in Chapter 5, I departed from the tagging of forms made by other scholars, from which I eliminated data points that

[3] At the purely formal level, one could see an analogue to the *sa*-aorist in the Greek A09 type (the "mixed aorist"), but the absence of sigmatic-thematic aorist formants in Avestan, or traces of such a formation elsewhere, and the complete absence of any cognate formations, renders such a possibility highly unlikely.

[4] *seṭ* ('with *i*') roots are Sanskrit roots having an abstract root-final segment that normally surfaces as *i* or *ī* between two consonants, Ø before a vowel, or which lengthens a preceding vowel when followed by a consonant. These effects result from phonological processes and sound changes relating to PIIr. */H/. *seṭ* roots are usually notated by a superscript *i*, e.g. $\sqrt{av^{i}}$ 'aid'; thus in the derivation of the 3.pl.*iṣ*-aor.act.ind. *āviṣuḥ*, the UR is given as /á-av$^{\text{i}}$-s-ur/.

[5] For instance, would it be better to treat the creation of the *siṣ* aorist as the outcome of a (complex and somewhat unconstrained, though conceivable) series of analogies, as proposed by Narten 1964, or as a case of morphological double marking, essentially resulting from the productivity of the *iṣ* aorist? As we will see, *siṣ*-aorists are restricted to roots in *-ā*, indicating that the formant /-siṣ-/ spread from some original members made to roots ending in *-ā*, however they came to be.

I judged incorrect or spurious, the risk of not having counted some potentially relevant data remains.

Data was input and arranged in essentially the same fashion as the Greek data, as represented in Table 5.2, though organized by Skt. root rather than lemma. In addition to the frequencies of each paradigmatic slot for each aorist, I also kept account of the present and perfect stems occurring to the same root in the RV, and their respective frequencies. Unlike for the Greek data, large-scale sorting, trimming, and re-classification of the original data set was not necessary.

However, the pertinence of two categories to aorist stems generally, and whether such forms ought properly to count towards the token frequency of an aorist stem, is necessary to consider. On the one hand are aorist participles, especially active participles, the place of which in the aorist systems has been recently evaluated by Lowe (2013); on the other hand are the so-called "*-si* imperatives".

Lowe's treatment of all forms that are morphologically susceptible to analysis as aorist participles finds that the majority of types and tokens should indeed be regarded as members of aorist paradigms. Lowe does, however, eliminate some forms from classification as aorist participles on four grounds: 1) the form is in fact the participle to a present or perfect stem; 2) the form, identical to an aorist middle participle, is in fact a "stative" (in the sense of Oettinger 1976) participle; 3) the form reflects a "Caland" adjective (i.e., an adjective built with the suffix *-ánt-* or *-āná-*), unconnected to a synchronic aorist stem; 4) the participle is a "non-participial nonce formation, with no necessary syntactic or semantic connection to the aorist in their formation." Quantitatively, to accept all of Lowe's decisions about what forms should genuinely be regarded as aorist participles does not substantially alter the composition of the data; the total number of tokens that would be removed (from root aorists, principally) is less than $100$; removing those forms from the data only marginally increases the value calculated for $\mathscr{P}$ (by less than $.001$) for root aorists, and does not impact the relative ranking of aorist types by $\mathscr{P}$ whatsoever. Nevertheless, I have tried to assess the data in light of Lowe's considerations, and applied them as follows: 1) the truly spurious aorist participle classifications made by Lubotsky identified by Lowe are thrown out; 2) because I disagree in principle with the notion that a separate "stative" category plays any role in Vedic morphosyntax, the classification of participles in this group as aorist participles is maintained; 3) in cases in which the possible "Caland" adjective stands alongside a finite aorist formation, the participle is kept;[6] 4) since what the morphological basis for forms in this group besides an aorist stem might be is difficult to see, I maintain participles belonging to this group as well. Again, however, I stress that the inclusion or exclusion of any or all members of these groups has no substantial import for the quantitative analysis of the Vedic aorist categories.

Of rather greater importance is the place of "*-si* imperatives" with respect to the aorist system. "*-si* imperatives" are a peculiar formation, occurring in the RV to $23$ roots, which are built with an apparent suffix *-si* to a verbal root, sometimes functioning as imperatives, at other times having future time reference (viz., as the Vedic subjunctive). Narten (1964: 45–

[6] Furthermore, Bozzone (2014) has argued that the entire "Caland" system rests originally on adjectival root aorists, and hence that these isolated participles should indeed be regarded as genuine evidence of root aorists.

6), Cardona (1965), and Szemerényi (1966) conclusively demonstrated that "*si*-imperatives" strongly co-occur with roots that show *s*-aorists or *sa*-aorists. In particular, a "-*si* imperative" is often attested to aorists that show a 3.sg.act.subj., which also usually lack a 2.sg.act.impv.; only to the roots $\sqrt{yudh}$ 'fight' and $\sqrt{pr̥}$ 'bring over, protect' do we find both a 2.sg.aor.act.impv. and a "-*si* imperative". From the functional point of view, that -*si* imperatives effectively fill the paradigmatic roles of the 2.sg.act.impv. and 2.sg.act.subj. is thus clear. The problem, on the formal side, is whether such forms are truly parsable as varieties of sigmatic aorist. The decision has some ramifications, because the total number of sigmatic aorist tokens ("-*si* imperatives" excluded) is small enough that the addition of the "-*si* imperatives" does substantially impact the calculation of productivity statistics; the principal effect is a notable decrease in the estimated value for $\mathscr{P}$ for *s*-aorists, though relative frequency rankings remain unaffected. By default, I report values including the "-*si* imperatives", and statistics on the Vedic aorists as a whole include the "-*si* imperative" tokens, unless otherwise specified.

One remaining problem is that certain paradigmatic forms are wholly ambiguous, formally, about what type of aorist they represent; this issue is most acute for 2.sg. and 3.sg. active indicative and injunctive forms, which may be interpreted as representative of either a) a root aorist to a *seṭ* root or b) an *iṣ*-aorist. For instance, $\sqrt{kram^{i}}$ 'step', clearly has a root aorist, as reflected in the 1.sg.act.ind. *akramīm* (RV 10.166.5c) and 3.pl.act.ind. *ákramur*; however, the 3.sg.act.ind. *akramīt*, attested 13×, could, in some instances, reflect a root derivation, but in others, an *iṣ* derivation – and what process, psychologically, a Vedic speaker used in the production of that particular form is impossible to say. Does *ákramīt* reflect derivation from /á-kram$^{i}$-t/ or /á-kram$^{i}$-(i)s-t/?

As a matter of fact, unambiguous instances of both root and *iṣ* aorists are found to just six *seṭ* roots, $\sqrt{kram^{i}}$ 'step', $\sqrt{gr̥bh^{i}}$ 'seize', $\sqrt{prath^{i}}$ 'spread', $\sqrt{van^{(i)}}$ 'win', $\sqrt{śnath^{i}}$ 'pierce', and $\sqrt{śram^{i}}$ 'be weary', and only $\sqrt{kram^{i}}$ and $\sqrt{gr̥bh^{i}}$ attest 2/3.sg.act.ind./inj. tokens (13× and 1×, respectively). Hence, the real impact at the level of tokens is negligible. The seeming root aorist to $\sqrt{gr̥bh^{i}}$, 1.sg.act.ind. *agrabham* is a hapax legomenon, and thus counting the 3.sg.act.ind. *agrabhīt* as an instance of a root rather than *iṣ*-aorist, the count of root aorist hapax legomena would be reduced by one.

## 6.3 Raw Results and Statistics

Having adjusted the token frequencies of participles as discussed above, when all causative aorists originally collected are removed from the dataset, some 328 distinct types remain; the basic frequency statistics are given in Table 6.2. In comparison to aorist frequencies in either of the Greek corpora examined in the preceding chapter, aorists in the RV are simply less diverse and less frequent. The aorist is evidently a less useful category in the RV than in Homeric Greek, even accounting for the overall differences in text size.[7] To my mind,

[7] $\chi^2$-tests evaluating the difference in number of aorist types with respect to number of tokens in the text suggests a non-significant difference between Homer and the New Testament ($\chi^2 = 3.563, p = 0.059$), but a highly significant difference between Homer and the RV (estimating the number of tokens in the RV at 165000, $\chi^2 = 106.8072, p < 0.0001$).

the interesting questions are the extent to which these substantial differences in frequency of usage of these cognate categories are to be attributed to differences in textual type (epic narrative poetry vs. religious lyric poetry), and the extent to which these differences are due to underlying differences in the Homeric and RVic grammatical systems.

| *V* | *N* | Mean *N* | Median *N* | Mode *N* | Max *N* | Std.Dev. |
|---|---|---|---|---|---|---|
| 328 | 4674 | 14.25 | 3 | 1 | 368 | 38.03955 |

Table 6.2: Basic Type and Token Statistics for Aorists in the *Ṛgveda*

| **AORIST TYPE** | *V* | *N* | $n_1 \sim \mathscr{P}*$ | $\mathscr{P}$ | $n_1/V$ | $S$ | $\mathscr{I}$ |
|---|---|---|---|---|---|---|---|
| Root | 117 | 2618 | 36 | .01375095 | .30769238 | 281.4676 | 2.426445 |
| Thematic | 45 | 868 | 14 | .01612903 | .3111111 | 120.9128 | 2.666667 |
| Reduplicated | 15 | 199 | 3 | .01507538 | .2 | 20.14282 | 1.258926 |
| Sigmatic | 151 | 1002 | 56 | .05588822 | .3708609 | 246.4216 | 1.631931 |
| *s*-Aorist | 75 | 711 | 21 | .02953586 | .28 | 121.5949 | 1.613333 |
| *iṣ*-Aorist | 66 | 248 | 32 | .1290323 | .4848485 | 114.8753 | 1.740535 |
| *sa*-Aorist | 7 | 29 | 3 | .1034483 | .4285714 | – | – |
| *siṣ*-Aorist | 3 | 15 | 0 | 0 | 0 | – | – |

Table 6.3: Productivity Statistics for Aorists in the *Ṛgveda*

Table 6.3, meanwhile, displays the usual set of productivity statistics for the non-causative aorist types, with the four flavors of sigmatic aorist individually broken out.[8] Again, the reader should bear in mind that these values, and especially the derived statistics $\mathscr{P}$ and $\mathscr{I}$, cannot be directly compared with the statistics calculated for other corpora. Overall, the ranking in productivity according to $\mathscr{P}$ is the same as found in Greek: root aorists are least productive, thematic and reduplicated aorists appear slightly more productive, and sigmatic aorists are decidedly productive. In particular, the *iṣ*-aorist shows an unmistakably high degree of productivity in comparison to all other varieties of aorist. Given the very small token samples of *sa*- and *siṣ*-aorists, how accurate the representation of those categories may be is questionable; in particular, the value of $\mathscr{P}$ at 0 for the *siṣ* aorists is slightly misleading, I think, given that two of the three types found in the RV are dis legomena. Moreover, if it is correct to say that *siṣ*-aorists are (originally, at least) a sort of *iṣ*-aorist, then their creation and extension is bound up with the very productivity of the *iṣ*-aorist. The *sa*- and *siṣ*-aorists will receive further consideration under 6.4.

[8] One other point of ambiguity which affects the calculations here is that some eight hapax legomena are 3.pl. forms (active *-an*, middle *-anta*) that may be either root or thematic aorists. Where Lubotsky explicitly labeled the form as either root or thematic, I have followed that classification, in all other cases, I have classified them as root aorists. Thus the $\mathscr{P}$ given for root and thematic aorists may, in reality, be somewhat lower and higher than reported there, though the relative ordering of the categories by that measure would remain unchanged regardless.

Although the relative productivity rankings achieved with the $\mathscr{P}$ statistic appear sensible, root aorists still contribute a substantial number of hapax legomena ($n_1 = 36$). When sigmatic aorists are divided into their constituent subcategories, the root aorists indeed show the largest number of hapax legomena among all categories. In accordance with the interpretation of $\mathscr{P}$ and $\mathscr{P}*$ offered in Hay and Baayen 2003 (see 3.4 above), the implication is that root aorists in the RV retain a good degree of usefulness, while being substantially less parsable than the varieties of sigmatic aorist. In short, unlike the athematic aorists of Homeric Greek, the root aorists of RVic Sanskrit appear to be a relatively robust category that has not yet suffered crippling depredations at the hands of more productive processes: even if there is not some means of creating novel root aorists in Vedic, a considerable underlying population of root aorists must be present. In fact, the estimated population sizes $S$ and levels of pragmatic potentiality $\mathscr{I}$, for both root and thematic aorists, point to the existence of considerably more types than occur in our actual sample. However, because the vocabulary size of those two categories (as measured by $V$) grows so slowly relative to the sigmatic aorists, a Vedic learner might fail to encounter many rare root or thematic aorist in his learning data. Consequently, this learner might instead build a sigmatic aorist, unaware of the root or thematic aorist present in the collective language.[9]

These points stand out in examining the vocabulary growth curves of our categories. Figure 6.1 displays the binomially interpolated VGCs for all four major aorist categories, up to the number of tokens attested for each category; the growth curves for hapax legomena are included as well. While the number of root aorist types appears to be increasing at a healthy rate, it pales in comparison to the rate of increase in types for sigmatic aorists.

Meanwhile, extrapolated growth curves up to $20000$ tokens for the root, thematic, and sigmatic categories given in Figure 6.2 show the very slow and steady increase in types to which the low $\mathscr{P}$ but higher $\mathscr{I}$ of root and thematic aorists point. Yet forms first appearing in this range would be rather rare types, occurring with a frequency of $\sim 1\times$ per million tokens.[10] The extrapolated growth curve for sigmatic aorists, on the other hand, exhibits a rapid increase in types, which nears its asymptote by $\sim 10000$ tokens. The larger point is that, in the lower frequency ranges, while Vedic speakers may have encountered considerable quantities of root and thematic aorist types, indicating that those categories were healthy and viable, they found that sigmatic aorist varieties were much more parsable, and probably used those to build genuine neologisms.

Nevertheless, root aorists show a surprising degree of seeming liveliness in comparison to other morphologically simplex categories in the RV. In particular, root presents (i.e., present stems that attach inflectional endings directly to a verbal root, and which show ablaut of the root) measure as substantially less productive than root aorists. $\mathscr{P}$ is considerably less for root presents than for root aorists; this data is given in Table 6.4. Even if one

[9] As Stephanie Jamison (p.c.) points out to me, a further piece of evidence supporting the productivity of the sigmatic aorist types in Vedic is the fact that the preterites in the Middle Indic languages consist principally of forms that correspond to the Sanskrit sigmatic aorist, though most of the types occurring in Middle Indic are nowhere attested in Sanskrit.

[10] Assuming a constant rate of token growth, to sample $20000$ root aorist tokens would require a corpus of approximately $1260000$ tokens of RVic genre text in total.

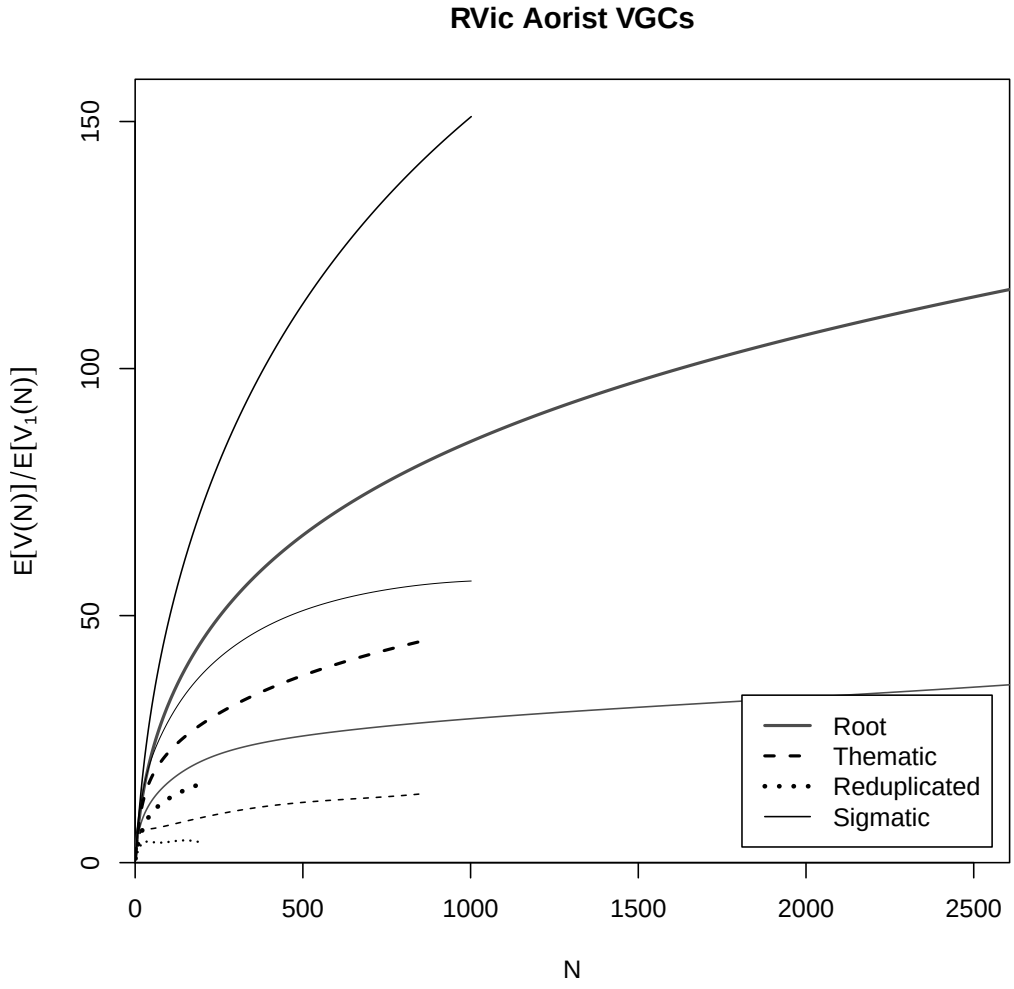



Figure 6.1: Vocabulary Growth Curves of Aorist Types in the RV

were to exlude the exceptionally frequent present to $\sqrt{as}$ 'be' ($1290$ tokens), which would bring the $N$ of root presents and root aorists into a comparable range, $\mathscr{P}$ and $\mathscr{P}*$ would measure as considerably less than for root aorists. It may be telling that Lubotsky 1998 marks most of the $10$ possible root present HL as nonce-formations; that is, Lubotsky doubts that those forms reflect genuine outputs of a natural grammar. I suspect that the quantitative differences between root aorists and root presents simply indicates historical differences in the population sizes of those two categories – root aorists long possessed many more types than root presents in the grammatical prehistory of PIE.

Finally, just as for Homer, a look at the composition of the most frequent aorist types is enlightening as well. Among the 30 most frequent types, listed in Table 6.5, not one type among the *iṣ*-, *sa*-, or *siṣ*-aorists occurs (the most frequent *iṣ*-aorist, *vadhiṣ*-, the *seṭ* quality

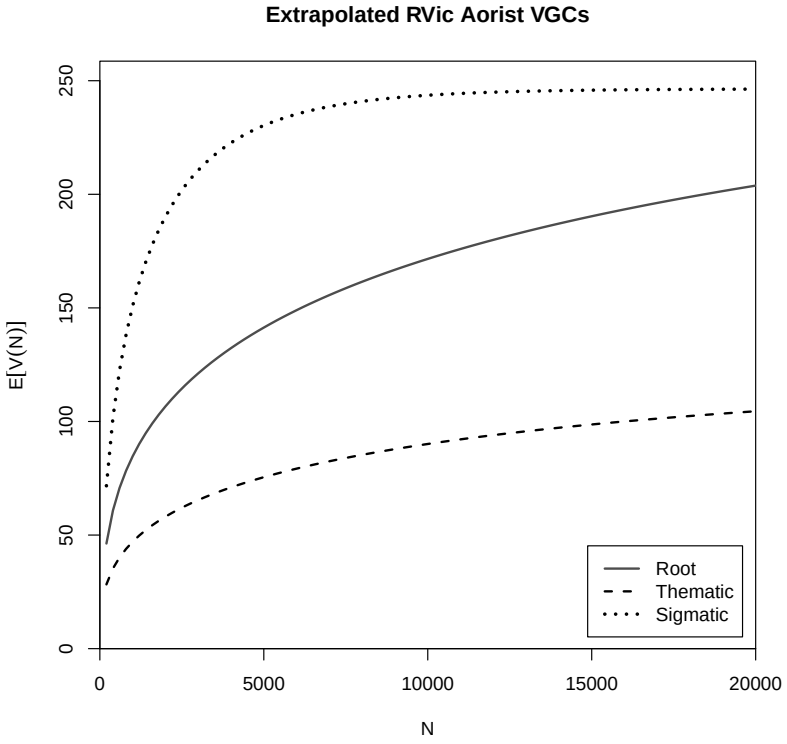



Figure 6.2: Vocabulary Growth Curves of Aorist Types in the RV

| $V$ | $N$ | $n_1 \sim \mathscr{P}*$ | $\mathscr{P}$ | $n_1/V$ | $S$ | $\mathscr{I}$ |
|---|---|---|---|---|---|---|
| 50 | 4217 | 10 | .002371354 | .2 | 73.12965 | 1.462593 |

Table 6.4: Productivity Statistics for Root Presents in the RV (including $\sqrt{as}$)

of whose root is uncertain, falls at rank 32). Despite the fact that sigmatic aorists together represent approximately 46% of all aorist types, they represent only 23% of the 30 most frequent types (7/30), and only one type in the 20 most frequent.

## 6.4 Analysis

A similar question to that which faced us in the analysis of the Greek aorist types confronts us here: why are the varieties of sigmatic aorist so productive in Vedic? For Greek, the answer ultimately seemed straightforward: first, derived present stems exceptionlessly form sigmatic aorists, and the morphophonological mappings between present stem and aorist stem were clearly much more reliable for sigmatic aorist patterns. The possibility of deriving novel athematic or thematic aorist stems, in accord with the measures $\mathscr{P}$ and $\mathscr{I}$ of those

| **ROOT** | Gloss | **AORIST STEM** | *N* | Absolute Rank |
|---|---|---|---|---|
| | | ROOT | | |
| √*gam* | 'come, go' | *gam-* | 368 | 1 |
| √*kṛ* | 'make, do' | *kar-* | 334 | 2 |
| √*bhū* | 'be, become' | *bhū-* | 281 | 3 |
| √*dhā* | 'put' | *dhā-* | 163 | 4 |
| √*¹naś* | 'attain' | *naś-* | 143 | 5 |
| √*sthā* | 'stand' | *sthā-* | 111 | 8 |
| √*śru* | 'listen' | *śro-* | 109 | 9 |
| √*gā* | 'come, go' | *gā-* | 92 | 10 |
| √*dā* | 'give' | *dā-* | 85 | 12 |
| √*²pā* | 'drink' | *pā-* | 72 | 16 |
| √*yuj* | 'yoke' | *yoj-* | 65 | 17 |
| √*idh* | 'kindle' | *edh-* | 56 | 19 |
| √*vṛ* | 'cover' | *var-* | 51 | 22 |
| √*yam* | 'reach' | *yam-* | 47 | 25 |
| | | THEMATIC | | |
| √*juṣ* | 'enjoy' | *juṣa-* | 138 | 6 |
| √*vid* | 'find' | *vida-* | 90 | 11 |
| √*bhū* | 'be, become' | *bhava-* | 78 | 13 |
| √*vidh* | 'worship' | *vidha-* | 74 | 14 |
| √*sad* | 'sit' | *sada-* | 73 | 15 |
| √*hū* | 'call' | *hūva-* | 63 | 18 |
| √*khyā* | 'see' | *khya-* | 44 | 26 |
| √*san*$^{i}$ | 'win' | *sana-* | 42 | 28 |
| | | REDUPLICATED | | |
| √*vac* | 'speak' | *voca-*[a] | 127 | 7 |
| | | SIGMATIC | | |
| √*rā* | 'give, bestow' | *rās-* | 55 | 20 |
| √*yaj* | 'sacrifice' | *yakṣ-* | 54 | 21 |
| √*yam* | 'reach' | *yaṃs-* | 49 | 23 |
| √*vah* | 'carry' | *vakṣ-* | 48 | 24 |
| √*nū* | 'roar, cry out' | *nūṣ-* | 43 | 27 |
| √*pṛ* | 'bring over, protect' | *parṣ-* | 41 | 29 |
| √*stu* | 'praise' | *stoṣ-* | 41 | 30 |

Table 6.5: 30 Highest-Frequency Aorist Stems in the *Ṛgveda*

[a] Whether this item is, in fact, synchronically transparent as a reduplicated formation is uncertain. Were one to treat it instead as a thematic aorist, the effect of adding this large number of tokens to the thematic aorist category would negatively impact the measure of productivity for that category. However, if the number of hapax legomena among the reduplicated aorists at 3 is correct, the effect of shifting *voca-* out of the reduplicated aorist would be to make this marginal category seem as productive as the sigmatic aorist ($3/72 = 0.04166667 = \mathscr{P}$). Below we will see that there is reason to doubt at least two of those reduplicated aorist hapax legomena as being genuine reduplicated aorists.

categories, seemed all but excluded. In Vedic, however, root and thematic aorists are not so evidently moribund as in Greek; the fact that the VGCs for those categories shown in Figures 6.1 and 6.2 above do not become asymptotic proves the point. I suggested above that the slow but steady growth curves of the Vedic root and thematic aorists would be consistent with a considerable population of low-frequency types existing in the language but not attested in the corpus; in virtue of their low frequency, they were liable to replacement by productive formations. Closer investigation is warranted, to see whether any genuine sources of productivity for the truly productive categories might exist.

In seeking to understand the productivity of Vedic sigmatic aorists, the crucial question is: from what sources do novel sigmatic aorists derive? Are they being built productively through a word-formation process combining root with sigmatic suffix? Are they being generated analogically from within previously existing aorist paradigms (e.g., by a mapping 3.sg. *-īt* → 1.sg. *-iṣam*, 3.pl. *-iṣan*, etc., thereby forming *iṣ*-aorists out of root aorists)? Or do other parts of the verbal paradigm accurately predict the corresponding type of aorist formation (perhaps most saliently, the type of present formation built to that verbal root)?

It is worth noting, at this point, the extent of competition in aorist formation present in the RV: nearly 30% of roots building any aorist build more than one type (94/328). In contrast, less than 10% (74/774) of all aorist types in Homer share a common root. This fact indicates that, whereas the Greek aorist system, as already attested in Homer, had already experienced the substantial expansion of sigmatic aorists, in the RV there remained more lively and ongoing competition between formations. I will explore some of the extent of this competition by examining the hapax legomena built to each of the Vedic aorist categories.

### 6.4.1 Profiles of Formational Types

#### 6.4.1.1 *iṣ*- and *s*-Aorists

It is helpful, from the outset, to examine the *iṣ*- and *s*-aorists, since these two subcategories contribute most to the number of aorist hapax legomena, and understanding the sources of the *iṣ*-aorists in particular is crucial. The first and most important point to be gleaned from Table 6.6 is that the vast majority of *iṣ*-aorist hapax legomena are built to *aniṭ* roots; only the forms built to the roots $\sqrt{aś^{i}}$ 'eat', $\sqrt{manth^{i}}$ 'shake, stir', $\sqrt{nū}$ 'roar', and $\sqrt{vand^{i}}$ 'extol' are *seṭ*. This fact indicates that novel *iṣ* aorists are emerging, in effect, from the suffixation of an independent morpheme /-iṣ-/. In 19 of the instances shown in the table, the *iṣ* aorist does not occur alongside any other aorist formation, which indicates that the *iṣ* aorist is indeed the productive source for aorist stems that may be genuine neologisms. The decisive conclusion must be that *iṣ*-aorists result, by and large, not from intraparadigmatic analogies affecting *seṭ* roots alone, but from the use of a free-standing morphological process.

Also notable are some instances of transparent derivation of the aorist stem from the present stem, rather than pure root. In particular, the 3.pl.act.ind. *ánindiṣuḥ* (incorrectly labeled as a reduplicated aorist by Lubotsky 1998: 814) stands out: this aorist stem *nindiṣ*- preserves the fossilized nasal infix of the present stem *ninda*-. Contrast this *iṣ*-aorist with the (also hapax) root aorist middle participle *nidānáḥ*, without the nasal infix. Given that

20 of these hapax legomena *iṣ*-aorists occur alongside present stems terminating in *-a* (i.e., Class I, IV, or VI presents), the operation of a morphophonological mapping [a-] → [iṣ-] / X_ may be at work. Some attraction of *iṣ* formations to specific paradigmatic slots may be inferred as well: seven of the hapax legomena in Table 6.6 are 3.sg.mid.ind./inj. forms that show the sequence *-iṣṭa*, whereas only one of the hapax legomena among the *s*-aorists (*áramsta*), given in Table 6.7, falls into this slot.

#### 6.4.1.2 *siṣ*-Aorists

The *siṣ*-aorist, in terms of types, is hardly an abundant category: only three roots exhibit the formation in the RV, and it appears to but a further eight roots in the entirety of Vedic literature. Moreover, there is a clear phonological restriction on the domain of *siṣ*-aorists: forms liable to analysis as such appear almost exclusively to roots in *-ā*; see Narten 1964: 70 for an enumeration of these forms.[11] This restriction indicates that, beyond some original core of *siṣ*-aorist roots, *siṣ* served as a genuine suffix /siṣ/ that could be used to build aorists, be it by derivation directly from a root, or by morphophonologically mapping from another surface form.

Although none of the instances of *siṣ* aorists in the RV are, strictly speaking, hapax legomena, the very low token frequency of the category as a whole, and the fact that two of the three types attested here are dis legomena, nevertheless suggests a high degree of parsability. I give all three attested forms in the RV in Table 6.8.

In terms of type frequency, the most common type of aorist built to roots in *-ā* is the root aorist, found in ten instances in the RV: $\sqrt{gā}$ 'come, go', $\sqrt{dā}$ 'give', $\sqrt{drā}$ 'run', $\sqrt{dhā}$ 'place', $\sqrt{pā}$ 'drink', $\sqrt{mā}$ 'measure', $\sqrt{rā}$ 'bestow', $\sqrt{śā}$ 'sharpen', $\sqrt{sā}$ 'bind', and $\sqrt{sthā}$ 'stand'. Hence, were aorist stems being generated by morphological processes mapping from the form of the abstract root to the aorist stem, to find the extension of *siṣ* aorists to other roots in *-ā* would be somewhat surprising, because the mapping [ā-] → [ā-] / C_ would be easily more reliable. $\sqrt{drā}$, though, is the only one of these root aorists to ever attest a *siṣ* aorist, and is in fact the least frequent of these root aorists, attesting both a hapax root aorist and a hapax *s*-aorist in the RV.

#### 6.4.1.3 *sa*-Aorists

Not only the hapax legomena (shown in Table 6.10), but indeed all roots to which *sa*-aorists are attested share the property of being built to roots that have a final segment that neutralizes to [k] before /s/, namely, *-ś*, *-j*, and *-h*. Furthermore, all *sa* aorists except for $\sqrt{mṛj}$ show either a Class I or a Class VI present, i.e., present stems that terminate in *-a*. By reconstructing the segment *h* in Vedic to its immediately preceding ancestor, $*\hat{\jmath}^{h}$ (= *[ɟ]), we can arrive to a neat minimal rule for describing the production of *sa* aorist forms from corresponding

[11] Macdonell (1916 [1993]: 166) asserts that *siṣ*-aorists are also found to roots in final nasals, but both are very late: *raṃsiṣam* appears as a mantra variant to the RVic *s*-aorist optative *rāsīya*, and *anaṃsīt* does not appear until the Vādhūla Sūtra. See Narten 1964: 73.

| ROOT | STEM | GLOSS | FORM | RV Citation | Other Aorist? |
|---|---|---|---|---|---|
| √²*akṣ* | *akṣiṣ-* | 'attain' | 3.pl.act.ind. *ā́kṣiṣuḥ* | 1.163.10d | – |
| √*aś*$^{i}$ | *aśiṣ-* | 'eat' | 3.sg.act.inj. *aśīt* | 10.87.17b | – |
| √*uh* | *uhiṣ* | 'consider, speak out' | 3.sg.mid.ind. *aúhiṣṭa* | 6.17.8c | – |
| *ūnay-*[a] | *ūnayiṣ-* | 'disappoint, fail (?)' | 2.sg.act.inj. *ūnayīḥ* | 1.53.3d | – |
| √*kṛp* | *krapiṣ-* | 'yearn' | 3.sg.mid.ind. *akrapiṣṭa* | 7.20.9b | root (1×) |
| √*ci* | *cayiṣ-* | 'gather' | 2.du.act.inj. *cayiṣṭam* | 6.67.8d | root (10×) |
| √*cud* | *codiṣ-* | 'impel' | 2.sg.act.inj. *codīḥ* | 1.63.4a | – |
| √¹*jambh* | *jambhiṣ-* | 'bite' | 3.sg.act.subj. *jambhiṣat* | 10.86.4c | – |
| √*das* | *dāsiṣ-* | 'waste away' | 3.sg.act.inj. *dāsīt* | 7.1.21d | – |
| √*dīv* | *daviṣ-* | 'play (dice)' | 1.sg.act.subj. *daviṣāṇi* | 10.34.5a | – |
| √¹*dhā* | *dhāyiṣ-* | 'put' | 2.sg.act.inj. *dhāyīḥ* | 1.147.5d | root, *s* (163×, 3×) |
| √*dhāv* | *dhāviṣ-* | 'stream' | 3.sg.mid.ind. *adhāviṣṭa* | 9.70.8b | – |
| √*nid* | *nindiṣ-* | 'insult' | 3.pl.act.ind. *ánindiṣuḥ* | 1.161.5b | root (1×) |
| √*nū* | *naviṣ-* | 'roar' | 3.sg.mid.ind. *anaviṣṭa* | 9.71.7b | redup. (2×) |
| √*pan*$^{i}$ | *paniṣ-* | 'admire' | 3.sg.mid.inj. *paniṣṭa* | 7.45.2c | – |
| √*bādh*$^{i}$ | *bādhiṣ-* | 'oppress' | 3.sg.mid.inj. *bādhiṣṭa* | 7.23.3c | – |
| √*budh* | *bodhiṣ-* | 'awaken' | 3.sg.act.subj. *bodhiṣat* | 2.16.7c | *s* (3×) |
| √*mad* | *madiṣ-* | 'exhilarate' | 3.pl.act.ind. *amādiṣuḥ* | 9.8.4c | *s* (37×) |
| √*manth*$^{i}$ | *manthiṣ-* | 'shake, stir' | 3.du.act.ind. *ámanthiṣṭām* | 3.23.2a | – |
| √*mṛṣ* | *marṣiṣ-* | 'neglect' | 2.sg.mid.inj. *marṣiṣṭhāḥ* | 1.71.10b | root (2×) |
| √*yam* | *yamiṣ-* | 'reach' | 3.sg.mid.inj. *yámiṣṭa* | 5.32.7b | *s* (51×) |
| √¹*yu* | *yaviṣ-* | 'bind' | 2.sg.act.inj. *yā́vīḥ* | 8.79.4c | *s* (17×) |
| √*rāj* | *rājiṣ-* | 'rule' | 3.pl.act.ind. *arājiṣuḥ* | 8.14.10c | – |
| √*ru* | *raviṣ-* | 'break' | 1.sg.act.inj. *rāviṣam* | 10.86.5c | – |
| √*vakṣ* | *vakṣiṣ-* | 'increase' | 2.sg.act.ind. *aúkṣīḥ* | 10.27.7a | – |
| √*vand*$^{i}$ | *vandiṣ-* | 'extol' | 1.pl.mid.opt. *vandiṣīmáhi* | 1.82.3b | – |
| √*viś* | *veśiṣ-* | 'enter, settle' | 3.sg.act.inj. *veśīt* | 8.60.20a | root, *s* (1×, 4×) |
| √²*viṣ* | *veṣiṣ-* | 'work' | 2.sg.act.subj. *veṣiṣaḥ* | 8.75.11b | – |
| √*śī* | *śayiṣ-* | 'lie' | 2.sg.mid.ind. *aśayiṣṭhāḥ* | 10.124.1d | – |
| √*sidh* | *sedhiṣ-* | 'repel' | 2.sg.act.inj. *sedhīḥ* | 10.27.20b | – |

Table 6.6: *iṣ*-aorist Hapax Legomena in the RV

[a] This item is not a root, but itself a derived denominative stem; *ūnayiṣ-* here appears to be the only example of any aorist built to such a derived stem.

| ROOT | STEM | GLOSS | FORM | RV Citation | Other Aorist? |
|---|---|---|---|---|---|
| √*kram*$^{i}$ | *kraṃs-* | ‘step’ | 3.sg.mid.subj. *kramsate* | 1.121.1d | root, *iṣ* (24×, 4×) |
| √*kṣi* | *kṣeṣ-* | ‘rule’ | 3.sg.act.subj. *kṣeṣat* | 6.3.1a | – |
| √$^{2}$*gṛ* | *gariṣ-* | ‘swallow’ | 3.sg.act.inj. *gārīt* | 5.40.7b | root (1×) |
| √*cit* | *cets-* | ‘perceive; appear’ | 3.sg.act.ind. *acait* | 6.44.7b | root (2×) |
| √*cyu* | *cyoṣ-* | ‘set in motion’ | 2.sg.mid.inj. *cyoṣṭhāḥ* | 10.173.2a | – |
| √*tan* | *tans-* | ‘stretch’ | 3.sg.act.ind. *atān* | 6.67.6d | root, thematic (7×, 3×) |
| √*tsar* | *tsarṣ-* | ‘steal’ | 3.sg.act.ind. *atsār* | 10.28.4c | – |
| √$^{3}$*dā* | *diṣ-* | ‘distribute’ | 1.sg.mid.opt. *diṣīya* | 2.33.5b | – |
| √*duh* | *dhukṣ-* | ‘rub, milk’ | 3.pl.mid.ind. *adhukṣata* | 9.110.8b | – |
| √*drā* | *drās-* | ‘run’ | 3.sg.act.subj. *drāsat* | 8.47.7b | root (1×) |
| √*dhū* | *dhūṣ-* | ‘shake’ | 3.pl.mid.ind. *adhūṣata* | 1.82.2b | redup. (2×) |
| √*dhvṛ* | *dhūrṣ-* | ‘injure’ | 3.pl.mid.ind. *dhūrṣata* | 5.12.5c | – |
| √*bhrī* | *bhreṣ-* | ‘injure’ | 3.sg.mid.subj. *bhreṣate* | 7.20.6a | – |
| √*mṛc* | *markṣ-* | ‘harm’ | 3.sg.mid.opt. (prec.) *mṛkṣīṣṭa* | 1.147.4d | – |
| √*ram* | *raṃs-* | ‘be calm’ | 3.sg.mid.ind. *áraṃsta* | 2.11.7d | – |
| √*vṛ* | *varṣ-* | ‘cover’ | 2.du.act.subj. *varṣathaḥ* | 8.5.21c | root (60×) |
| √*vṛj* | *vṛkṣ-* | ‘twist’ | 1.sg.mid.inj. *vṛkṣi* | 1.27.13d | root (16×) |
| √*śī* | *śeṣ-* | ‘lie’ | 3.pl.act.subj. *śéṣan* | 1.174.4a | *iṣ* (1×) |
| √*śri* | *śreṣ-* | ‘affix, resort’ | 1.pl.act.subj. *śreṣāma* | 4.43.1d | root (22×) |
| √*spṛ* | *sparṣ-* | ‘rescue, save’ | 1.sg.act.ind. *áspārṣam* | 10.161.2d | root (8×) |
| √*syand* | *syants-* | ‘flow | 3.sg.act.ind. *asyān* | 9.89.1a | redup. (10×) |
| √*svan*$^{i}$ | *svans-* | ‘sound’ | 3.sg.act.inj. *svānīt* | 2.4.6b | root (1×) |

Table 6.7: *s*-aorist Hapax Legomena in the RV

| ROOT | STEM | GLOSS | $N$ | RV Citation | Other Aorist? |
|---|---|---|---|---|---|
| √$^{2}$*gā* | *gāsiṣ-* | ‘sing’ | 2 | 8.1.7d, 8.81.5a | *s*, *iṣ* (3×, 1×) |
| √$^{1}$*yā* | *yāsiṣ-* | ‘go’ | 11 | … | *s* (9×) |
| √$^{2}$*yā* | *yāsiṣ-* | ‘implore’ | 2 | 1.18.6c, 4.1.4b | *s* (5×) |

Table 6.8: *siṣ*-aorist Stems in the RV

| ROOT | STEM | GLOSS | Form | RV Citation | Other Aorist? |
|---|---|---|---|---|---|
| $\sqrt{kruś}$ | *krukṣ-* | 'cry out' | 3.sg.act.ind. *ákrukṣat* | 10.146.4d | – |
| $\sqrt{ruh}$ | *rukṣ-* | 'ascend; grow' | 3.sg.act.ind. *árukṣat* | 10.67.10b | thematic (31×) |
| $\sqrt{vṛh}$ | *vṛkṣ-* | 'tear' | 1.sg.act.ind *avṛkṣam* | 10.159.5c | – |

Table 6.9: *sa*-aorist Hapax Legomena in the RV

| ROOT | STEM | GLOSS | FORM | RV Citation | Other Aorist? |
|---|---|---|---|---|---|
| $\sqrt{taṃs}$ | *tataṃsa-* | 'pull away, tug' | 2.du.act.ind. *átatamṣatam* | 1.120.7b | – |
| $\sqrt{pū}$ | *pupo-* | 'purify' | 3.sg.act.ind. *ápupot* | 3.26.8a | *iṣ* (3×) |
| $\sqrt{śvit}$ | *śiśvita-* | 'be bright' | 3.sg.act.ind. *áśviśvitat* | 8.5.1b | root, *s* (2×, 4×) |

Table 6.10: (Non-causative) Reduplicated Aorist Hapax Legomena in the RV

present stems: $C_{[+high]}$a- → kṣa-. Given that all of the hapax legomena here occur in the latest part of the RV, we might attribute their production to the active workings of two separate subrules into which the proto-rule [$C_{[+high]}$a-] → [kṣa-] decomposes, once **ʃʰ* becomes Ved. *h*:

$C_{[+high]}$a- → [kṣa-] (on the basis of pres. *mṛja-* and *mṛśa-* : aor. *mṛkṣa-* and *mṛkṣa-*) generates *krukṣa-* from pres. *kruśa-*.

ha- → [kṣa-] (on the basis of pres. *guha-* and *duha-*) generates *rukṣa-* and *vṛkṣa-* from pres. *ruha-* and *vṛha-*.

#### 6.4.1.4 Reduplicated Aorists

In terms of types, tokens, and hapax legomena, reduplicated aorists make up the smallest category outside of the *sa-* and *siṣ-*aorist subcategories. Two of the hapax legomena here are thematic, while one is not; two are 3.sg.act.ind. forms, and both of those occur alongside other infrequent aorists. I see no notable unities or patterns characterizing either these hapax legomena or the class as a whole, as would be expected of a largely unproductive category. Indeed, the category may be even more unproductive, given that both the stems *tataṃsa-* and *pupo-* may, in fact, ultimately reflect perfect subjunctive stems that have been reanalyzed an thematic indicative stems; see further Kümmel 2000: s.v.v. *taṃs-* and *pav*$^{i}$ for further discussion of this possibility. The very fact the overall quantitative picture of the reduplicated aorist suggests a feeble category may, then, give one good reason to believe that these two items ultimately reflect stems of the (surely much more productive) Vedic reduplicated perfect.

| ROOT | STEM | GLOSS | FORM | RV Citation | Other Aorist? |
|---|---|---|---|---|---|
| $\sqrt{^2as}$ | *asa-* | 'throw' | 3.pl.act.inj. *asan* | 4.3.11a | – |
| $\sqrt{kṛt}$ | *kṛta-* | 'cut, split' | 2.sg.act.ind. *ákṛtaḥ* | 1.63.4d | – |
| *kṛpaṇ-* | *kṛpaṇa-* | 'long for, desire' | 3.pl.mid.inj. *kṛpáṇanta*[a] | 10.74.3b | – |
| $\sqrt{gṛdh}$ | *gṛdha-* | 'be greedy' | 3.sg.act.ind. *gṛ́dhat* | 10.34.4b | – |
| $\sqrt{gṛh}$ | *gṛha-* | 'complain' | 1.pl.mid.inj. *gṛhāmahi* | 8.21.16b | – |
| $\sqrt{gṝ}$ | *gura-* | 'sing, welcome' | 2.sg.mid.impv. *gurasva* | 3.52.2b | – |
| $\sqrt{dam^i}$ | *dama-* | 'control, subdue' | 2.sg.act.inj. *dáṉaḥ*[b] | 1.174.2a | – |
| $\sqrt{piś}$ | *piśa-* | 'adorn' | 2.sg.act.impv. *piśa* | 7.18.2c | root (1×) |
| $\sqrt{puṣ}$ | *puṣa-* | 'thrive' | 1.pl.act.opt. *puṣema* | 10.128.1b | – |
| $\sqrt{vṛdh}$ | *vṛdha-* | 'grow' | 3.sg.act.ind. *avṛdhat* | 3.38.2c | root (4×) |
| $\sqrt{sridh}$ | *sridha-* | 'fail' | 3.sg.act.inj. *sridhat* | 7.34.17b | – |

Table 6.11: Thematic Aorist Hapax Legomena in the RV

[a] The status of this form as an aorist is doubtful – it may rather be a Class I impf.inj., given that it appears parallel to the clear 3.pl.mid.impf.inj. *kṛṇávanta* (to $\sqrt{kṛ}$ 'make, do') in the immediately preceding verse.

[b] The fact that this form shows an unexpected *-n-*, rather than *-m-*, is troubling, and gives one reason to doubt the status of this form. Stephanie Jamison (p.c.) suggests to me that it is a badly mutilated imperfect of a Class IX present, thus **damnāḥ*, which would be supported by the existence of the *-āyá-* stem *damāyá-* (< Transponat **demn̥h₂i̯é-*).

#### 6.4.1.5 Thematic Aorists

Cardona (1960) systematically argues for the thesis that thematic aorists in Vedic consistently resulted from the mis-parsing of 3.pl.act./inj. root forms: the sequence *-an* was interpreted as *-a-n*, and then used to build thematic aorist paradigms, on the analogy of thematic imperfects, where 3.pl. *-an* would regularly map to 3.sg. *-at*, 2.sg. *-aḥ*, etc. Although only two of these hapax legomena occur alongside root aorists, more generally, $14/45$ thematic aorists attested in the RV occur alongside a root aorist, by my reckoning; Cardona's thesis, from a quantitative point of view, thus seems possible.

#### 6.4.1.6 Root Aorists

On the whole, I see no systematic regularities that bind together the set of root aorist hapax legomena listed in Table 6.12. In fact, the one possible uniting characteristic is that these aorists belong to genuinely low-frequency verbal roots: the median token frequency of the most frequent present stem corresponding to these root aorist hapax legomena is $3$ whereas the median token frequency of the present stems to *iṣ* aorist and *s* aorist hapax legomena is $10.5$ and $9.5$ respectively. This difference indicates that the HL among sigmatic types occur to better attested and more well established verbs, whereas the HL among root aorists consist, in the main, of more poorly attested verbs. Such a distinction is consistent with the interpretation that the RVic root aorist HL are more likely to reflect very rare lexemes (i.e., be relics) than productive creations.

Nevertheless, we do encounter a few peculiarities, such as the seeming 3.sg.act.subj. *tan-*

*drat* built to a stem *tandr-*. *tandrat* itself is usually regarded as an "Augenblicksbildung" (so Hoffmann 1967: 240, with Gotō 1987: 159 and Kümmel 2005: 321 following him), in fact occurring alongside two other possible root aorist subjunctives *tamat* and *śramat* in the same pāda. While the latter two look like possibly well-formed root aorists, their appearance with the strange *tandrat*, raises suspicion. Given the low productivity of the root aorist in any case, dismissing these three forms would not change the overall portrait of the category.

### 6.4.2 Correlations of Present Types and Aorist Types

Precisely because the Vedic verbal system presents a sometimes bewildering array of different verbal formations constructed on the basis of a single root, it is difficult to state or determine with any degree of certainty what reliable relationships hold between different verbal stems without an infeasible amount of data entry. Although I have alluded to some possibly valid reliable mappings from present stems to aorist stems that may hold good in Vedic, the task of constructing a working model that predicts the form of aorist stems on the basis of other verbal stems belonging to the same root must be a task left to the future. In lieu of such a complete learning simulation and predictive model, I will comment here on some co-occurrence correlations that hold between present stems and aorist stems in the *Ṛgveda*.

The present stem is the likely core element (base) of Vedic individual verbal systems built to given roots for the following reasons: 1) it is usually the most frequent member of the system, and 2) it is the member with the fewest phonological neutralizations on the root, at least in the case of thematic presents. Just among the $328$ aorist stems used in the investigations here, the most common present stem is typically more frequent than the aorist $(243/328)$; in the same data, the most common present stem is more common than the perfect stem in an almost equivalent number of cases $(241/328)$.

Most striking in the data here is the tendency of sigmatic aorists to occur alongside Class I presents (i.e., present stems with a suffix *-a*, with accent on the root syllable), and of the varieties of nasal infix present stems to occur alongside root aorists. Co-occurrence frequencies are given in the Table 6.13 below. The contrast in frequency of nasal presents between root and sigmatic aorists is almost certainly not due to chance ($\chi^2 = 15.2215, p < .0001$), while the contrast in frequency of Class I presents between root and sigmatic aorists looks unlikely to be due to chance ($\chi^2 = 3.5722, p = .05875$). The marked correlation observed here supports the general thesis of Strunk (1967), that the nasal infix was a principal means of deriving present stems from existing root aorist stems, as against the criticism of, e.g., Beekes (1969: 279). Assuming the productivity of the Class I presents in Sanskrit, the productivity of varieties of sigmatic aorist could then follow from this existing correlation: perhaps Class I presents tend to spawn sigmatic aorists in much the same fashion that the Greek denominative and deverbative formations tend to spawn sigmatic aorists.

| ROOT | STEM | GLOSS | FORM | RV Citation | Other Aorist? |
|---|---|---|---|---|---|
| √*kṛp* | *kṛp-* | 'yearn' | 3.pl.act.ind. *akṛpan* | 4.2.18c | *iṣ* (1×) |
| √*gṛbh*$^{i}$ | *grabh-* | 'seize' | 1.sg.act.ind. *agrabham* | 1.191.13c | *iṣ* (5×) |
| √$^{1}$*gṝ* | *gūr-* | 'welcome, sing' | 3.sg.mid.inj. *gūrta* | 1.173.2c | thematic (1×) |
| √$^{2}$*gṝ* | *gar-* | 'swallow' | 3.pl.act.subj. *garan* | 1.158.5a | *iṣ* (1×) |
| √*chid* | *ched-* | 'cut' | 1.pl.act.inj. *chedma* | 1.109.3a | – |
| *tandr-* | *tandr-* | 'fatigue' | 3.sg.act.subj. *tandrat* | 2.30.7a | – |
| √*tam*$^{i}$ | *tam-* | 'tire' | 3.sg.act.subj. *tamat* | 2.30.7a | – |
| √*tṛd* | *tard-* | 'bore' | 3.sg.act.subj. *tardat* | 6.17.1a | – |
| √*drā* | *drā-* | 'run' | 3.pl.act.impv. *drāntu* | 10.85.32d | *s* (1×) |
| √*druh* | *druh-* | 'be hostile' | 3.pl.act.inj. *druhan* | 1.5.10a | – |
| √*dhvaṃs* | *dvas-* | 'create smoke/dust' | 3.pl.act.inj. *dvasán* | 8.55.5c | – |
| √*dhvan*$^{i}$ | *dhvan-* | 'smoke' | 3.sg.act.ind. *dhvanīt* | 8.6.13a | – |
| √*nid* | *ned-* | 'insult' | nom.pl.mid.part. *nidānắḥ* | 4.5.12d | redup. (1×) |
| √*nud* | *nud-* | 'push' | 2.sg.mid.inj. *nutthắḥ* | 6.17.5d | – |
| √*piś* | *piś-* | 'adorn' | nom.pl.mid.part. *piśānắḥ* | 7.57.3c | thematic (1×) |
| √$^{1}$*pī* | *pī-* | 'swell' | nom.sg.mid.part. *píyānaḥ* | 1.79.3a | – |
| √*pruṣ* | *pruṣ-* | 'sprinkle' | 1.sg.act.subj. *pruṣā* | 10.77.1a | – |
| √*mud* | *mud-* | 'rejoice, delight' | 1.pl.mid.opt. *mudīmahi* | 8.1.14d | – |
| √*myakṣ* | *myakṣ-* | 'be situated, affix' | 3.sg.act.ind. *ámyak* | 1.169.3a | – |
| √$^{1}$*yu* | *yav-* | 'bind' | 3.pl.mid.inj. *yavanta* | 5.2.5a | *iṣ* (1×) |
| √*rabh* | *rabh-* | 'take hold' | 3.sg.mid.ind *árabdha* | 10.8.3a | – |
| √*ric* | *ric-* | 'leave' | 2.sg.mid.inj. *rikthāḥ* | 3.6.2b | *s* (5×) |
| √*riṣ* | *reṣ-* | 'be hurt' | 3.sg.act.subj. *reṣat* | 7.20.6a | thematic (34×) |
| √*vij* | *vij-* | 'agitate, stir up' | 3.sg.mid.inj. *vikta* | 1.162.15b | – |
| √*vip* | *vip-* | 'tremble' | nom.sg.mid.part. *vipānáḥ* | 8.6.29c | – |
| √*viś* | *viś-* | 'enter, settle' | 3.pl.mid.ind. *áviśran* | 8.27.12d | *s*, *iṣ* (4×, 1×) |
| √*śram*$^{i}$ | *śram-* | 'be weary' | 3.sg.act.subj. *śramat* | 2.30.7a | *iṣ* (2×) |
| √*sagh* | *sagh-* | 'support' | 3.sg.act.subj. *sághat* | 1.57.4c | – |
| √*sac* | *sac-* | 'follow' | nom.sg.mid.part. *sacānáḥ* | 6.20.2d | *s* (6×) |
| √*sañj* | *saj-* | 'hang' | 3.sg.mid.ind. *asakta* | 1.33.3a | – |
| √*skand* | *skand-* | 'leap' | 3.sg.act.inj. *skán* | 10.61.7a | – |
| √*stubh* | *stubh-* | 'praise' | nom.sg.mid.part. *stubhānáḥ* | 4.3.12c | – |
| √*spaś* | *spaś-* | 'spy' | 3.sg.mid.ind. *áspaṣṭa* | 1.10.2b | – |
| √*spūrdh* | *spūrdh-* | 'contend' | 3.pl.act.inj. *spūrdhán* | 6.67.9a | – |
| √*sphṛ* | *sphar-* | 'move suddenly' | 2.sg.act.inj. *sparīḥ* | 6.61.14b | – |
| √*svan*$^{i}$ | *svan-* | 'sound' | 3.sg.act.ind. *ásvanīt* | 4.27.3a | *s* (1×) |

Table 6.12: Root aorist Hapax Legomena in the RV

| Aorist Type | $V$ | Class I Present | Nasal Present (Classes V, VII, IX) |
|---|---|---|---|
| Root | 117 | 31 | 30 |
| Thematic | 45 | 9 | 4 |
| Sigmatic | 151 | 66 | 8 |

Table 6.13: Frequencies of Vedic Aorist Types vis-à-vis Class I and Nasal Presents

## 6.5 Comparison to Greek Aorists and Implications for the Reconstruction of Proto-Indo-European Aorist Formants

With precise measures of productivity for the Vedic aorist categories now in hand, we are now in the position to attempt the comparison with the cognate categories of Greek as reflected by the data from Homer presented in Chapter 5. Already under section 6.2, I noted that the ranking in productivity of the Vedic aorist categories essentially corresponds to the ranking found in Greek. For instance, exactly as is theoretically expected and diachronically traceable, root aorists are not productive, either in the RV or Homer – no systematic and learnable means of deriving root aorists exists in either language, and so we have every reason to think that the same applies to their proto-language.[12] Likewise, the marginal productivity of thematic aorists in both Homer and the RV suggests that the category did not tend ever to expand by truly productive word-formation processes.

Just as was done in order to assess the relative productivity of aorist categories between Homer and the New Testament in section 5.5, appropriate reductions in sample $N$ to the Homeric aorist categories, and estimation of the corresponding growth curves through binomial interpolation permits direct comparison of categories across corpora. Figures 6.3, 6.4, and 6.5 plot the type and hapax VGCs for the categories of root, thematic[13], and sigmatic aorists for Homer and the RV.

The difference between Greek and Sanskrit in the realm of the root aorists is truly stark: whereas the category in the RV remains useful and well-populated, the category in Homer is clearly moribund. Indeed, the growth curve of hapax legomena in Homer is the only hapax curve among those shown in Figures 6.3–5 that clearly begins to fall towards $0$. In contrast, although the sigmatic aorist was evidently the most productive among the Vedic aorist varieties, the category in Greek is substantially more productive: note that, within the token sample size of $1002$, the total number of sigmatic aorist types in the RV barely comes to exceed the number of estimated hapax legomena for the Greek. Once again, the outsized productivity of the Greek sigmatic aorist depends upon its potentiation by apparently

[12] More properly, since the only languages under comparison here are Greek and Vedic Sanskrit, one must speak of the last common stage of the Greek and Indo-Iranian branches (and whatever other Indo-European subgroups one believes necessarily to fall into that unity).

[13] While reduplicated aorists in Vedic have been treated separately up to this point, I have collapsed them with the Vedic thematic aorist data at this point, because the Greek thematic aorist data includes reduplicated aorists. On both the Greek and Vedic side, tokens of IE */ u̯e-u̯kʷ-e/o-/ 'spoke' make up a large proportion of the tokens in the reduplicated aorist/Ao6 categories.

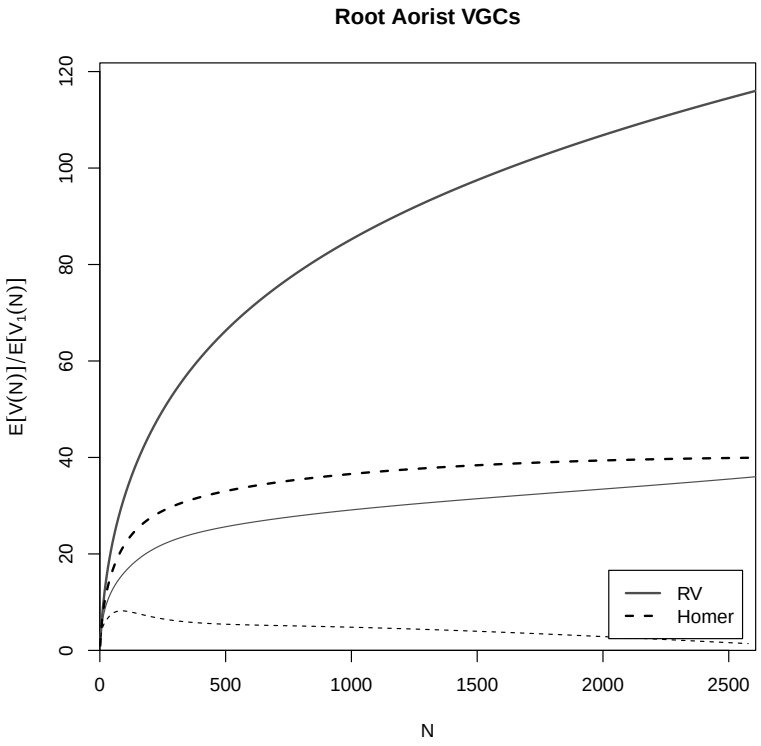



Figure 6.3: Interpolated Growth Curves of Root Aorists in the RV and Homer

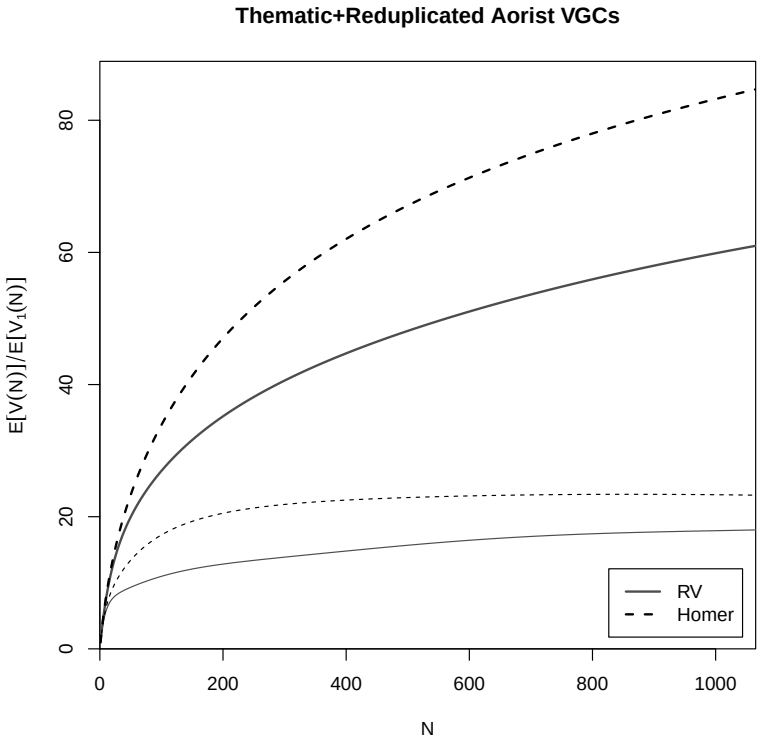



Figure 6.4: Interpolated Growth Curves of Thematic (combined with reduplicated) Aorists in the RV and Homer

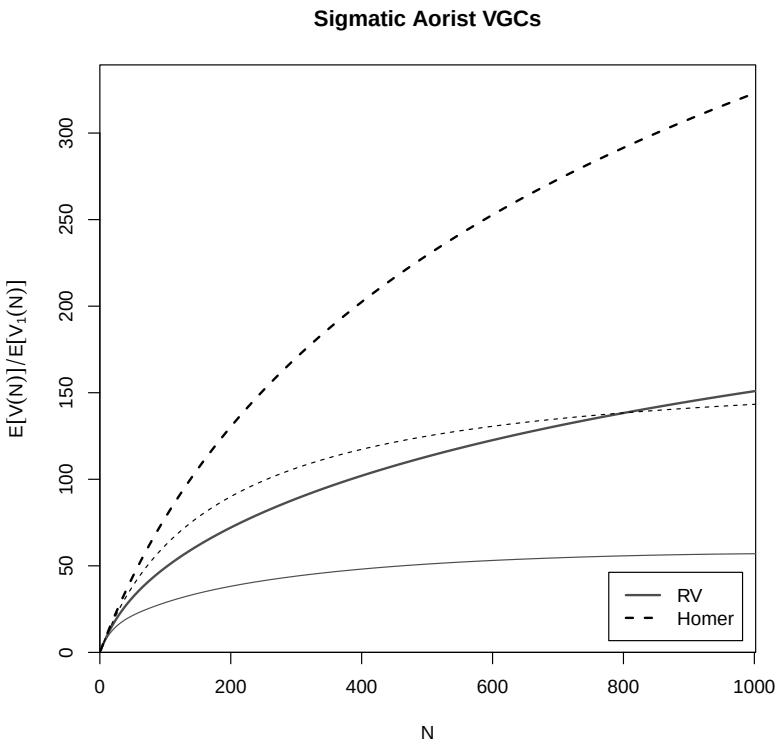



Figure 6.5: Interpolated Growth Curves of Sigmatic Aorists in the RV and Homer

productive denominal and deverbal derivation, as discussed at 5.4.1.3. Non-primary verbal formations, such as denominatives (e.g., *apasyáti* 'works' ← n. *apás-* 'work'), seem to have substantially fewer types in the RV than do denominatives in Homer, though I believe that the productivity of Vedic denominative verbs is probably rather high (I sense that the category knows a substantial number of hapax legomena).[14] More relevant to the present issue, however, is the fact that early Vedic denominatives do not seem to build aorists, but rather, only presents. The only clear example of an aorist to a denominative stem in the RV appears to be the 2.sg.inj. *ūnayīḥ* 'you disappointed (?)'; telling though, is the fact that its aorist is an *iṣ* aorist. Although differing substantially in details, the aorist categories of the RV and Homer agree in having a generally productive sigmatic aorist, and a generally unproductive to moribund root aorist.

Likewise, in both languages, thematic (taken together with reduplicated thematic) aorists occupy a position of marginal productivity. In a surprising number of details, the RVic and Homeric thematic aorists exhibit quantitative similarities. For instance, the number of types $V$ belonging to thematic aorists makes up a similar proportion of all aorist types in both the RV and Homer ($\sim$ 15–18% $= \frac{117}{774}, \frac{60}{328}$). Even the total population estimate $S$ of the category in the two languages is remarkably close: $\sim$ 126 for Homer, $\sim$ 120 for the RV, and those estimated populations are not distant from the number of types actually found in the Homeric corpus, 117. Finally, as the hapax growth curves in Figure 6.4 show, the actual degree of productivity $\mathscr{P}$ is not dissimilar either. I find this quantitative similarity across the two languages striking.

Overall, this attempt to quantitatively confirm the the status of the aorist categories in Greek and Vedic does not substantially alter the existing portrait and conception of those categories for the last common ancestor of Greek and Sanskrit. The low productivity of the root aorist means that philologically secure root aorist forms may be safely interpreted as archaisms, even without good comparative evidence. Precisely the opposite holds true for the sigmatic aorists: very good comparative evidence, extending beyond Greek and Indo-Iranian, should be adduced to reconstruct specific sigmatic aorist forms. Yet, the shared fact of great productivity suggests that the common ancestor of Greek and Vedic also shared some germ of productivity. This, too, is no surprise to the Indo-Europeanist – Latin, Celtic languages, and Slavic languages all attest a substantial number of perfective and preterite forms built with morphology cognate to the Greek and Indic sigmatic aorists, but with relatively few cognate stems.[15]

[14] The figure given by Tucker (1988) of "approximately 100 formations" built with denominative *-yá-* in the RV seems correct to me. The list of secondary verbal formations found in Homer listed in Tucker 1990: 429–505 exceeds 300 types.

[15] For a general survey of the place of sigmatic aorist formations in Indo-European languages generally, and their particular historical developments within the Celtic subgroup, see Watkins 1962. Likewise, on the history of the Italic sigmatic aorist, see Meiser 2003: Ch. 7; for the place of the sigmatic aorist in Slavic, see now Ackermann 2015.

## 6.6 Brief Reflections on Chapters 5 and 6

Taken together, I believe that the most important result to follow from the preceding two chapters is the finding that quantitative corpus-based measures of productivity seem to obtain sensible results, even when applied to these small corpora of less than $200000$ tokens. The fact that these methods seem to pass a general "sanity check" on easy cases should instill confidence that their outputs should be taken seriously when applied to less obvious cases. In addition, the predictions that these methods can generate about the subsequent history of the language may create interesting research questions in themselves. For instance, imagine that the New Testament data in Chapter 5 had shown an increase in productivity among thematic aorists, or even a failure of the non-productive categories to exhibit appreciable declines in class membership; such conditions contrary to the straightforward prediction made by the Homeric data would call for an account. My hope is that these quantitative methods can be successfully applied to problems where they might bring genuine insight; the most fruitful potential ground would in fact be problems that go beyond questions of morphological productivity *per se*, but which can be explained, at least in part, by recourse to a category's productivity. Section 7.4 in the following chapter attempts to do just this, by considering $\mathscr{P}$ as a measure of morphological parsability, and consequent implications for accentuation where morphological structure is relevant.

Similarly, the fact that the MGL simulation carried out in 5.4.2 not only captured the fundamental unlearnability of most Greek root and thematic aorists, and captured the extension of some sigmatic sub-patterns that accord with the historical record, should be taken as another successful empirical test of the model on a relatively easy case. The formalization of a few MGL-style rules above also could explain the creation of the *sa*-aorist hapax legomena attested in the RV; such rules help to describe the phonological conditioning with greater perspicuity than a four-part analogy. Chapter 8 will show that the attention to detail that the MGL affords us can, moreover, be of aid in refining and testing the limits of specific hypotheses about diachronic phonological and morphological changes.

# CHAPTER 7

# Productivity Effects in Ancient Greek and Sanskrit Accentuation

The principles that underlie Vedic Sanskrit and Ancient Greek word-level prosody ("accentuation") remain among the most difficult-to-comprehend aspects of their respective phonological systems. For the modern student or scholar of these two languages, accentuation presents a thicket of dense underbrush, in which even the most reliable patterns admit of some apparent exceptions. To the linguist, the absence of a self-evident surface systematicity raises the question of learnability: beyond brute memorization, how did Vedic and Greek speakers possibility acquire the accentuation of each lexeme, and on what basis could they compute the accentuation of a novel lexeme? The objective of this chapter will be to explore some factors that act as determinants of Vedic and Greek accentuation. More specifically, I will examine the role that morphological structure plays not just at the level of morphological categories, but individual lexemes. Thus, the problem is to try to bring out how the productivity of morphological processes and the frequency of specific lexemes may be reflected in the prosody of that lexeme.

This entire enterprise is predicated upon work carried out in the preceding chapters. In Chapter 3, we saw psycholinguistic evidence to the effect that the perceived morphological structure of a stem or whole word is likely determined by many different lexical frequency relations. In Chapters 5 and 6, I showed that the corpus-based methods for the measurement of productivity introduced in Chapter 2 appear to yield sensible results in simple cases. The quantitative assessment of productivity, as measurable by $\mathscr{P}$, $\mathscr{P}*$, and $\mathscr{I}$ (cf. 2.2.2), permits us to estimate how "parsable" the forms of a given category generally are. Granting that accentuation in Greek and Vedic is indeed sensitive to morphological structure, differences in parsability could conceivably lead to differences in surface accentuation.

Once again, the principal evidence will derive from the corpus frequencies that the Homeric Epics and the *Ṛgveda* provide to us, though examples from Attic-Ionic generally or other Vedic texts will appear here.[1] The focus here will lie with nominal categories, because Greek verbs always exhibit a "default" phonological accent, and for independent (perhaps historically related) reasons, verbal accent in Sanskrit is often not directly attested.[2] Section 7.1 first

[1] While I accept that the accentual data of the *Ṛgveda* is essentially reliable, one should note that the addition of prosodic information to the texts of Homer substantially post-dates their commission to writing. Hence, there exists the real possibility that the accentuation for at least some Homeric forms is not its original or true accentuation, but the accentuation assigned to it by a later scholar. See Chantraine 1958: Ch. XV for a brief overview of the Homeric accentual tradition and some particularities of accentuation in the manuscripts of the text.

[2] Specifically, all forms of finite verbs in main clauses that do not stand in some prosodically prominent

introduces the essential concepts and facts that are necessary to describe accentuation in Sanskrit and Greek, largely following the account in Kiparsky 2010. Section 7.2 treats some problematic details of Kiparsky's model; I there attempt to constrain the number of conceptual entities needed for a system of lexical accentuation (principally by reference to the work of Revithiadou 1999), and discuss one case in Vedic Sanskrit where explicit attention to *lack* of internal morphological structure helps to resolve instances of undergeneration in Kiparsky's system. With a working general account of Greek and Sanskrit accentuation, I then review how frequency effects have previously been related to Greek accentuation by Probert (Probert 2006b, Probert 2006a). 7.4 then takes up the examination of accentuation in several nominal categories in Greek and Vedic. 7.5 concludes with discussion of the further work needed to achieve a model of Greek and Sanskrit accentuation that can aspire to explanatory adequacy.

## 7.1 A Brief Description of Greek and Sanskrit Accentuation

### 7.1.1 Some History and Terminology

Traditionally, the placement of the principal word accent (i.e., determination of the syllable with greatest prominence) in Sanskrit and Greek is described as "free": the languages show relatively weak (Greek) or seemingly no (Sanskrit) phonological restrictions on which syllable may licitly bear the word accent (the ICTUS) within any given sequence of syllables.[3] In other words, the Sanskrit word accent is totally unbounded (i.e., in a string of $X$ syllables, the peak prominence may appear on any syllable, regardless of how many there may be), while the Greek word accent is free to surface within the rightmost three syllables. Although descriptions of the "Law of Limitation" and the tendency for certain word classes (especially, all verbs) to show "recessive" accentuation in Greek go back at least to the 19th century, the principles according to which a given word might show a given accent have remained obscure.[4] For both Greek and Sanskrit, in both the historical indigenous and Western 19th and 20th century grammatical traditions, more than observationally accurate accounts of the basic facts (e.g., all past passive participles built with the suffix *-tá-* in Sanskrit always bear the word accent on the syllable of that morpheme) did not exist until the latter half of the 20th century.

In short, both Sanskrit and Greek possess so-called "lexical accent" systems, in which underlying lexically specified accentual properties of constituent morphemes compete for faithful realization in a prosodic word. Where no underlying lexical accents to which the

position are transmitted without accentuation.

[3] For reasons described below, I will adopt the term "ictus" to refer to the property that marks a syllable as the most prominent in a prosodic word. This term is slightly unfortunate, because it may sometimes connote a stress accent, and hence is not ideal to attach as a term to the prosodies of Ancient Greek and Sanskrit, which both possessed pitch accents. Nevertheless, I will use "ictus" because it covers a necessary distinction, and I do not wish to introduce a new term.

[4] For a highly reliable and accessible description of Ancient Greek accentuation, including some ancient testimonia, see Probert 2003.

phonology must be faithful are present in a morphological assemblage, a purely phonological means of calculating the word-level prominence is employed. To my knowledge, Kiparsky (1973) was the first to clearly describe the behavior of the Sanskrit and Greek word accent in such terms, though the basic theoretical conception that Kiparsky inherited came out of work on Slavic accentology (cf. Illich-Svitych 1963, translation in Illich-Svitych 1979, and Dybo 1968), and similar accentual phenomena had already been described for Cupeño (Hill and Hill 1968). Following Kiparsky (2010), I will refer to the word-level prominence that has a particular phonetic realization as the "ictus", and employ "accent" to mean an abstract phonological property involved in the determination of the ictus.

Two crucial theoretical assumptions must underlie the reckoning of a language's word-level prosodic system as a "lexical accent" system. First is the assumption that prosodic information may constitute part of the lexical entry of roots, affixes, stems, and even full word forms, in just the same way that the particular phonemic content of given morphological entities is lexically stored. This basic assumption seems unproblematic. Slightly trickier is the assumption that phonological grammars can learn to be faithful to lexical accents, and have some means of adjudicating between multiple lexical accents. In 7.1.2 below, I show that a constraint set that fulfills the following basic demands is readily formalizable:

- each prosodic word contains at most and at minimum one syllable with peak prominence (i.e., one ictus per prosodic word).
- in case the morphology provides no lexical accents, phonological constraints should consistently predict the same type of ictus.
- in case the morphology provides multiple lexical accents, the phonology must have a consistent means of deciding to which lexical accent to be faithful.

The remainder of this section then gives a cursory overview of other issues related to Greek and Sanskrit accentuation.

### 7.1.2 Basics: Vedic and Greek Lexical and Default Accents

#### 7.1.2.1 The Sanskrit BAP

As a point of departure, I take the the algorithm for Sanskrit accent assignment proposed by Kiparsky, termed the Basic Accentuation Principle, as he formulates it in Kiparsky 2010: 144:

(11) Basic Accentuation Principle (BAP): Erase all accents but the leftmost one, and put an accent on the leftmost syllable of an unaccented domain.[5]

[5] Kiparsky (Forthcoming) is somewhat clearer, requires less unpacking, and makes explicit reference to tonal properties:

(1) Basic Accentuation Principle (BAP):

a. The leftmost High toned syllable/mora of a domain retains High tone, the others get Low tone.

Kiparsky treats Sanskrit accentuation within a derivational framework, with the application and cyclicity of rules decided and constrained by the addition of morphemes (i.e., Lexical Phonology; cf. Kiparsky 1982b). The BAP is thus a post-lexical combination of constraint (i.e., don't allow more than one ictus) and rule (select the leftmost accent or insert an accent) that sees the accents ultimately output by the preceding lexical cycles. In the event that multiple accents are output to the post-lexical phonology, the leftmost accent is selected as the ictus; in the event that no accents are output to the post-lexical phonology, an accent is assigned to the leftmost syllable, and surfaces as the ictus. The BAP evidently fulfills the three requirements given at the end of the preceding subsection: it ensures that each word has one and only one ictus, and selects which accent should become an ictus. If we take for granted a metrical theory of accent (Halle and Vergnaud 1987; both Revithiadou (1999) and Alderete (2001), describing lexical accent systems generally, and Kim (2002), working on Indo-European accentuation, do so), the behaviors set forth by the BAP can be obtained with four constraints given in 12, ranked as in 13:[6]

(12) Constraints underlying the BAP:

a. CULMINATIVITY: Assign a violation for each prosodic word that lacks an ictus (existence requirement) or which has more than one ictus (uniqueness requirement).

b. MAX-IO(Accent): Assign a violation for every accent present in the input that does not correspond to an ictus in the output.

c. DEP-IO(Accent): Assign a violation for every ictus present in the output that does not correspond to an accent in the input.

d. ALIGN-L(I(ctus), Pr(osodic) Word): Assign a violation for every syllable (the domain to which an ictus can apply) that intervenes between the ictus and the left edge of the word.

(13) Ranking: CULMINATIVITY $\gg$ MAX-IO(Accent) $\gg$ DEP-IO(Accent), ALIGN-L(I, PrWord)

CULMINATIVITY crucially dominates both MAX-IO(Accent) and DEP-IO(Accent), thereby ensuring that accents will be inserted and deleted as necessary to satisfy it. MAX-IO(Accent) crucially dominates ALIGN-L(I, PrWord); otherwise, underlying accents would be irrelevant, and the ictus would always appear on the the leftmost syllable. DEP-IO(Accent) is not crucially ranked with respect to MAX-IO(Accent) or ALIGN-L(I, PrWord); any insertion of accents is gratuitous when an underlying accent is available, and inserting an accent other than on the leftmost syllable would gratuitously violate ALIGN-L(I, PrWord).[7]

b. If there is no High toned syllable/mora, put High tone on the leftmost syllable/mora.

[6] In general, with Revithiadou (1999: 193), I believe that we can consider morphemes in languages with lexical accent systems to be learned with metrical information as part of the morpheme's subcategorization matrix.

[7] For simplicity, at this point, I assume that accents cannot reassociate to other syllables than the syllable by which they are hosted in the input, and hence any difference in accents between input and output involves violation of MAX-IO(Accent) and/or DEP-IO(Accent).

The basic typology can be seen in the combination of a monosyllabic root with an inflectional ending, with one instance each of accentedness or unaccentedness. Here, I take the Sanskrit roots /pad-/ 'foot' and /gáv-/ 'cow', to be lexically unaccented and accented, and the inflectional morphemes /-as/ 'nom.pl.' and /-ás/ 'gen./abl.sg.' to be lexically unaccented and accented, respectively.[8] Indeed, the two inflectional endings /-as/ and /-ás/ are, in Sanskrit, formally identical apart from their accentual properties. These combinations and their resolutions are presented in 14–17. The root morphemes also exhibit ablaut; I ignore this fact for the moment and simply assign it to the UR.

(14) /pād-as/, 'foot' nom.pl.

| pād-as | | Culminativity | Max-IO(Accent) | Dep-IO(Accent) | Align-L(I, PrWord) |
|---|---|---|---|---|---|
| a. | pādas | ∗! | | | |
| b. ☞ | pā́das | | | ∗ | |
| c. | pādás | | | ∗ | ∗! |
| d. | pā́dás | ∗! | | ∗∗ | |

(15) /pad-ás/, 'foot', gen.sg.

| pad-ás | | Culminativity | Max-IO(Accent) | Dep-IO(Accent) | Align-L(I, PrWord) |
|---|---|---|---|---|---|
| a. | padas | ∗! | ∗ | | |
| b. | pádas | | ∗! | ∗ | |
| c. ☞ | padás | | | | ∗ |
| d. | pádás | ∗! | | ∗ | |

(16) /gā́v-as/, 'cow', nom.pl.

[8] In point of fact, clear cases of accented roots such as /gáv-/ seem to be rare in both Sanskrit and Greek. As we shall see, cases of persistent paradigmatic ictus on the syllable corresponding to the root are often attributable to other factors, specifically, Output-Output faithfulness to the position of the ictus in the nominative or accusative singular.

(17) /gáv-ás/, ‘cow’, gen.sg.

| gắv-as | | | Culminativity | Max-IO(Accent) | Dep-IO(Accent) | Align-L(I, PrWord) |
|---|---|---|---|---|---|---|
| a. | | gāvas | ∗! | ∗ | | |
| b. | ☞ | gắvas | | | | |
| c. | | gāvás | | ∗! | ∗ | ∗ |
| d. | | gắvás | ∗! | | ∗ | |

| gáv-ás | | | Culminativity | Max-IO(Accent) | Dep-IO(Accent) | Align-L(I, PrWord) |
|---|---|---|---|---|---|---|
| a. | | gavas | ∗! | | ∗∗ | |
| b. | ☞ | gávas | | ∗ | | |
| c. | | gavás | | ∗ | | ∗! |
| d. | | gávás | ∗! | | | |

Comparison of the paradigms for the cognate forms of ‘foot’ in Sanskrit and Greek, given in 18, shows an essentially identical pattern of accentual movement between the root syllable and inflectional endings; only the accentuation of the acc.pl. disagrees. Moreover, the analysis seems to follow from the same basic principles: where we analyze no underlying accent, the ictus (high tone) appears to the left, but where an underlying accent is available, it receives the ictus.[9]

(18) Skt. /pad-/ ‘foot’ and Gk. /pod-/ ‘foot’

| **Case** | /pắd-/ | | | | /pod-/ | | | |
|---|---|---|---|---|---|---|---|---|
| | Sg. | | Pl. | | Sg. | | Pl. | |
| | UR | SR | UR | SR | UR | SR | UR | SR |
| Nom. | /pād-s/ | *pắt* | /pād-as/ | *pắdas* | /pod-s/ | *pó:s* | /pod-es/ | *pódes* |
| Acc. | /pād-am/ | *pắdam* | /pad-ńs/ | *padás* | /pod-n/ | *póda* | /pod-ns/ | *pódas* |
| Inst. | /pad-ắ/ | *padắ* | /pad-bhís/ | *padbhíḥ* | – | – | – | – |
| Dat. | /pad-áy/ | *padé* | /pad-bhyás/ | *padbhyáḥ* | /pod-í/ | *podí* | /pod-sí/ | *posí* |
| Abl. | /pad-ás/ | *padáḥ* | /pad-bhyás/ | *padbhyáḥ* | – | – | – | – |
| Gen. | /pad-ás/ | *padáḥ* | /pad-ắm/ | *padắm* | /pod-ós/ | *podós* | /pod-ô:n/ | *podô:n* |
| Loc. | /pad-í/ | *padí* | /pad-sú/ | *patsú* | – | – | – | – |

#### 7.1.2.2 The Greek Law of Limitation and Recessive Accent

As mentioned above, while there are no constraints on the surface position of the ictus in Vedic (e.g., one encounters forms such as *ánūnavarcās* [nom.sg., of Agni] ‘having a flawless

[9] Note that the Gk. dative suffix /-í/ is cognate to the Sanskrit locative suffix /-í/.

sheen' or *paripanthínam* 'having a path around'), Attic-Ionic Greek tightly constrains the licit domain within which a high tone may surface: a high tone may not appear more than one vocalic mora to the left of the rightmost bimoraic foot in a word.[10] This constraint is the Law of Limitation. The essence of the Law of Limitation should be both to drive the default "recessive" accent (illustrated in 20) and to constrain the domain of lexical accents; the same fundamental motivation ought to keep both the phonological and lexical ictus within the same prosodic territory.

Steriade (2014) considers the Greek high tone to be the ictus, and to be equivalent to a stress, and regulates the licit domain of a high tone through two *Lapse constraints:

(19) Steriade (2014)'s constraints deriving the Law of Limitation:
   a. *ExtendedLapseR: assign one violation * for every right edge of a word containing more than two stressless syllables.
   b. *LapseR$\bar{\sigma}$: assign one violation * for every right edge of a word containing two stressless syllables at the right edge, if the final is heavy.

The effect of these two *Lapse constraints is entirely descriptively adequate: they ensure that, at latest, a high tone occurs on the third syllable from the right edge of the word. However, because these conditions refer purely to syllables, an analysis in these terms misses the compelling generalization of the "recessive" accent in terms of foot structure. According to the analysis of Golston (1990) (cf. Sauzet 1989, Probert 2010: 2–5, Gunkel 2010: 25–7), Attic-Ionic builds moraic trochees from right to left, and assigns the principal word ictus (marked with an * below) around the rightmost foot in the word. Note also that word-final consonants are extrametrical. Final consonant extrametricality renders final -VC syllables light, which presumably leaves them unparsed when a heavy syllable occurs immediately to their left (as in nom.sg. ἄνθρωπος [(án.)(t$^{h}$rɔ:.)po⟨s⟩] 'man'). (20) gives examples of several recessively accented forms.[11]

(20) Attic-Ionic Default Ictus Assignment
   a. nom.sg. (πο.λυ.)(άν.)(*θρακ)⟨ς⟩ [(po.lu.)(án.)(*t$^{h}$rak⟨s⟩)] 'having much coal'
   b. nom.sg. (ἄν.)(*θρω.)πο⟨ς⟩ [(án.)(*t$^{h}$rɔ:.)po⟨s⟩] 'man'
   c. acc.pl. (ἀν.)(θρώ.)(*που)⟨ς⟩ [(an.)(t$^{h}$rɔ́:.)(*po:⟨s⟩)].
   d. 1.sg.impf.mid.ind. ἐ.(δυ.νά.)(*μη)⟨ν⟩ [e.(du.ná.)(*mɛ:⟨n⟩)] 'I was able'
   e. 1.pl.impf.mid.ind. ἐ.(δυ.νά.)(*με.θα) [e.(du.ná.)(*me.t$^{h}$a)] 'we were able'
   f. **False**: 1.sg.impf.mid.ind. $^{X}$ἐ.(δύ.να.)(με.θα) [e.(dú.na.)(me.t$^{h}$a)]. The high tone stands more than one vocalic mora to the left of the rightmost foot.

[10] All coda segments in Greek are weight-bearing, but absolute word-final consonants are extrametrical; thus the sequence -VC# counts as light and -VCC# counts as heavy.

[11] I employ the traditional values of the Greek acute and circumflex symbols in writing Greek forms: an acute marks a high tone on a short-voweled syllable or a high tone on the second mora of a long-voweled syllable (thus, a rising tone), while a circumflex marks a high tone on the first mora of a long-voweled syllable (thus, a falling tone).

What we call the ictus, for Attic-Ionic Greek, should not be considered merely the association of a high tone to a given syllable, as in Sanskrit, but rather the association of a high-low tonal contour to a sequence of two vocalic morae. An ictus defined as a particular tonal sequence rather than a single tone-to-syllable mapping will be subject to somewhat different concerns. In general, the Attic-Ionic ictus strongly recalls the way in which Japanese lexical accents are realized, namely, as a high-to-low contour (cf. Pierrehumbert and Beckman 1988). In words in which the lexical accent aligns with the rightmost syllable of a lexical item, this high-low contour cannot be realized unless an additional syllable is added. Thus, in (21), the tonal distinction between the forms 'fence' /ka**ki**/ and 'persimmon' /kaki/ is not evident without the addition of a clitic to the right (**boldface** indicates the lexical specification of the high-toned part of the ictus). The default tonal pattern applies in (21).c., where all moras have high tone following the initial low tone.

(21) Accentual Minimal Pairs in Tōkyō Japanese

a. /**ka**ki/ 'oyster' → [kákì] : /**ka**ki-ga/ 'oyster' nom. → [kákìgà]

b. /ka**ki**/ 'fence' → [kàkí] : /ka**ki**-ga/ 'fence' nom. → [kàkígà]

c. /kaki/ 'persimmon' → [kàkí] : /kaki-ga/ 'persimmon' nom. → [kàkígá]

The same applies to Attic-Ionic: for instance, lexically accented ἀδελφός [adelp$^{h}$ós] 'brother' defines where the ictus must begin, but the low part of the contour cannot surface unless an additional syllable to the right appears within the domain of the prosodic word, e.g., ἀδελφός τις [adelp$^{h}$ós tis] 'some brother'. Unlike Japanese, because the default ictus placement of Attic-Ionic never results in a high tone on the final syllable, a tonal pattern arising from a lexical accent on a final syllable is never conflated with the default tonal pattern, as is the case for the Japanese forms in (21).b. and c.

The Attic-Ionic realization of the ictus is almost surely innovative vis-à-vis Sanskrit, and probably innovative with respect to Proto-Greek. The very limited evidence that we have on the accentuation of Doric Greek suggests the existence of a similar Law of Limitation, but we regularly find the high tone in Doric one mora to the right of where it surfaces in Attic-Ionic. Compare Att.-Ion. 1.sg.aor.act.ind. ἔλαβον [élabon] 'I took' vs. Dor. ἐλάβον [elábon] (cf. Buck 1955: 85, Probert 2003: 160–2). If we assume that the default ictus assignment in Doric is sensitive to the rightmost foot, just as in Attic-Ionic, but resulted in the assignment of a high tone, rather than low tone, to the rightmost foot, this difference falls out easily: /elabon/ is parsed as ἐ(*λαβο⟨ν⟩) [e.(*labo⟨n⟩)] in both dialects, but a different phonetic realization results.

I propose that the Law of Limitation reflects the need to align part of the ictus with part of the phonological word's head foot. Such a constraint tries to obtain close tonal and metrical alignment. Since the Attic-Ionic ictus divides over two morae, the ideal satisfaction of metrical and tonal alignment is achieved when the left and right edges of the ictus are aligned with the left and right edges of a bimoraic head foot. Such alignment alone, however, predicts that recessively accented forms in Attic-Ionic would exhibit the high tone within the rightmost foot. For instance, Attic-Ionic would show $^{\text{X}}$[e.(lábo⟨n⟩)], rather than the correct [é.(labo⟨n⟩)] given above. Given that adequate alignment can be achieved provided that at

least one part of the ictus falls within the head foot, other constraints seem to be at work to force the high tone leftward in Attic-Ionic. Perhaps the best account is to say that the low-toned part of the ictus strives for alignment with the left edge of rightmost foot.

Such leftward shift of the high tone is also apparent in the paradigms of certain stems with a lexical accent on a long vowel – the pattern here was under the traditional label of the "σωτῆρα [sɔ:tɛ̂:ra] Rule". The descriptive observation is that stems with a long vowel at the right edge of the stem show acute accentuation when word-final (e.g., nom.sg. σωτήρ [sɔ:tɛ́:r] 'savior') or when followed by a heavy syllable (e.g., gen.pl. σωτήρων [sɔ:tɛ́:rɔ:n]), but circumflex accentuation when followed by a light syllable (e.g., acc.sg. σωτῆρα [sɔ:tɛ̂:ra], hence the name of the rule). Since the ictus in nom.sg. σωτήρ [sɔ:tɛ́:r] is non-recessive, it must be lexical, thus /sɔ:tɛ́:r-/. Gunkel (2013) plausibly argues that the σωτῆρα [sɔ:tɛ̂:ra] Rule can be motivated as the avoidance of tonal crowding at the right edge of a word. Given that an ictus consists of a tonal contour, its phonetic realization is optimized and clarified the further back the the peak (high tone) stands from the right edge of the word. Alternations such as acc.sg. [sɔ:tɛ̂:ra] : gen.pl. [sɔ:tɛ́:rɔ:n] are licit granted that the constraint driving the σωτῆρα Rule outranks a constraint militating against movement of an underlying accent, and that part of the ictus always remains within bounds of the head foot. Since the leftward movement of the high tone is motivated by phonetic optimization of a tonal contour, it follows that languages in which the ictus is just a single tone, no phonetically motivated constraints will attempt to place more distance between the high tone and the right edge of the word to ensure that the tonal pattern of the ictus is clearly realized. This prediction appears to be borne out by the fact that Doric does not exhibit the σωτῆρα Rule (cf. Probert 2006b: 71; but see complicating details in Hinge 2006: 124–8).

For present purposes, the precise interaction of numerous constraints to obtain the patterns of "default" recessive accentuation and lexical accentuation is not crucial. The reader need simply bear in mind that lexical accents are constrained by the Law of Limitation and modified by the σωτῆρα Rule. I use the following four constraints, in which the Law of Limitation, Sōtêra Rule, and Recessive Accent must be understood as cover constraints for the interaction of constraints on foot structure, prosodic alignment, and the licit position of a high tone.

(22) a. Law of Limitation: a high tone must not occur farther to the left in a prosodic word than one vocalic mora to the left of the rightmost foot.

b. Sōtêra Rule: the configuration μ́.μ# (where all μ are vocalic morae) is disallowed.

c. Ident-IO(Accent): an underlying accent linked to a mora in the input must be realized as a high tone on that mora in the output.

d. Recessive Accent: place the ictus as far leftward in a prosodic word as possible.

(23) **Ranking**: Law of Limitation, Sōtêra Rule ≫ Ident-IO(Accent) ≫ Recessive Accent.

Thus, only the Law of Limitation and Sōtêra Rule may modify the position of an accent in the input, while accents in the input are preferred to the Recessive Accent. At 7.1.5, we will

see that some further modifications may be necessary to explain why accentual mobility, in both Greek and Vedic, is relatively uncommon in nominal paradigms.

### 7.1.3 Typology of Accentual Properties

In the foregoing, I have followed a considerable body of other scholarship in holding that Greek and Vedic possess default phonological ictus assignment mechanisms driven by alignment constraints. We also saw that some degree of faithfulness to lexically specified accents is a further necessary component. Given that morphemes can be encoded with prosodic information, the question becomes: what is the range of prosodic properties needed? Kiparsky (2010, Forthcoming) operates with six possible combinations of three features controlling accent location (ACCENTED, PRE-ACCENTING, and UNACCENTED) and two features controlling relations between accents (DOMINANT and RECESSIVE); their combinatorial possibilities, with a claimed instance of each in Vedic, are given in Table 7.1.[12]

| | RECESSIVE | DOMINANT |
|---|---|---|
| ACCENTED | 3.pl.pres.act.ind./-ánti/ | past passive participle /-tá-/ |
| PRE-ACCENTING | subjunctive /-ˊăˊ-/ | superlative /-ˊiṣṭhaˊ-/ |
| UNACCENTED | part.mid. /-(m)āna-/ | agent /-tar-/ |

Table 7.1: Typology of Accentual Properties

In a rough sense, dominant morphemes are said to "erase" the other underlying accents of other morphemes, and impose their own. Recessive morphemes, on the other hand, do not impact the accentual properties of their neighbors. For instance, the Vedic root $\sqrt{takṣ}$ 'fashion', synchronically appears to be accented /tákṣ/, given the 3.pl.pres. *tákṣati* (1× RV); contrast *adánti* 'they eat', *sánti* 'they are', which suggest a suffix /-ánti/, accented but non-dominant, like gen.sg. /-ás/.[13] The PPP, however, is *taṣṭá-* 'fashioned' ← /tákṣ-tá-/: the accent of /-tá-/ would appear to win over the accent of /tákṣ-/, despite the phonological preference to align the ictus with the left edge of the word.

The distinction between dominant and recessive morphemes seems to interact in crucial respects with the morphology. In particular, Kiparsky notes that dominant morphemes never fall outside of recessive morphemes, and that, in the framework of Lexical Phonology, dominant morphemes are level 1 morphemes. More generally, there appears to be a decided correlation between dominant morphemes and derivational morphemes; Steriade (1988a) goes so far as to assert that all derivational morphemes of Greek are dominant (cf. also Probert 2006b: 145–6). Steriade is correct to observe that all derivational suffixes, when attached to either a root that seems to be accented, or another derived form with the ictus on

[12] I say "claimed instance of each" because fully verifying the correct accentual properties of a given morpheme is not always straightforward, and I do not necessarily agree with all of the analytical choices made by Kiparsky.

[13] The form of the inflectional ending in *tákṣati*, *-ati* may be due to a post-ictic vowel syncope, and surface *-ati* then reflects underlying /-nti/. See Kiparsky Forthcoming: 20.

its derivational suffix, either attract the ictus to themselves (presumably dominant accented suffixes) or induce recessive accentuation (presumably dominant unaccented suffixes). See examples in (24) below.

(24) Accentual Dominance in Greek Derivation

a. ἀστήρ [asté:r] 'star', dat.sg. ἀστέρι [astéri] implies /as-tér-í/ (cf. [podô:n] ← /pod-ô:n/): /as-tér-ísko-s/ → ἀστερίσκος [asterískos] 'little star'. The accent of /-ísko-/ takes precedence over /-tér-/

b. ἀληθής [alɛ:tʰɛ:s] 'true', dat.pl. ἀληθήσι implies /alɛ:tʰ-é:-sí/: /alɛ:tʰ-é:-ia/ → ἀλήθεια [alé:tʰeia] 'truth', with recessive accent. The unaccentedness of /-ia-/ takes precedence over the accent of /-é:-/.[14]

The underlying forces that drive dominance effects are of crucial relevance to the question of relations between productivity and accentuation, and hence will be more closely examined and motivated under 7.2.

### 7.1.4 Ablaut?

Indo-Europeanists have long accepted some deep connection between accentuation and patterns of vowel alternation; Hirt (1900) tried to argue in detail, and Schindler (1975), for instance, accepted that, at some early stage of (Pre-)PIE, [-high] vowels (i.e., */e, o, a/) were subject to syncope when not under the ictus. The surface ictus of Greek and Sanskrit clearly has relatively little direct connection to patterns of vowel alternation. For Greek, instances of ablaut probably need to be handled as either lexicalized forms or through morphological constructions. In Sanskrit, at least one variety of ablaut applies with sufficient regularity as perhaps to rise to the level of phonology: the deletion of underlying /a/ or /ā/ preceding an accented morpheme. Kiparsky (2010: 145) defines a simple rule:

(25) ZERO GRADE
/a, ā/ → Ø/_ morpheme$_{[+accent]}$
An underlying /a/ or /ā/ is deleted before an underlyingly accented morpheme.

For instance, Kiparsky ascribes the alternation between [tar] and [tr] in acc.sg. /bhŕ̥-tar-am/ → [bhŕ̥taram] 'brother' and inst.sg. /bhŕ̥-tar-ā́/ → [bhŕ̥trā] to the operation of this rule, which refers to underlying accents, not the surface ictus. The veracity and reliability of this rule is relevant for the detailed investigation of Sanskrit accentuation, since it might indicate covert traces of an underlying accented morpheme.

However, for a model of Optimality Theory without levels or strata (contrary to Kiparsky's approach), a rule of this sort is worrisome because it requires an intermediate level of representation in the phonology between input and output – since accents that trigger ZERO GRADE are not (always) present in the output, the faithfulness violation induced by vowel

[14] But note Hom. ἀληθείᾱ, Ionic ἀληθείη. Since the rightmost vowel is long rather than short, thus the recessive accent establishes the placement of the high tone on the [í].

deletion cannot be motivated by any markedness constraint. The application of ZERO GRADE in inst.sg. /bhŕ̥-tar-ā́/ 'brother' would presumably apply in the word-level phonology, after the attachment of inflectional morphology; the intermediate UR /bhŕ̥-tr-ā́/ is then passed to the post-lexical phonology, which then selects the leftmost accent as the ictus, thereby rendering the application of ZERO GRADE in this form opaque. Specifically, the ictus assignment process often counterbleeds the ZERO GRADE rule. By assigning the ictus to the first syllable in *bhŕ̥trā*, the environment for ZERO GRADE (an accented morpheme following an /a/ or /ā/) is destroyed. Examining the paradigm acc.sg. *bhŕ̥tāram*, inst.sg. *bhŕ̥trā* alone would fail to provide the learner with evidence for the property that conditions the application of ZERO GRADE.

Zero grade would appear most consistently to apply before accented inflectional endings in *r*- and *n*-stems. Greek, meanwhile, preserves only some traces of zero-grade ablaut in the same stem classes: kinship *r*-stems show zero grade of the suffix /-ter-/ (e.g., dat.pl. πατράσι [patrási] < *[pəh$_2$tr̥si]), but no such ablaut in productively derived agent nouns with the same suffix (e.g., nom.sg. δωτήρ [dɔːtéːr] 'giver', gen.sg. δωτῆρος [dɔːtêːros]).

At the same time, Sanskrit shows a sort of vowel insertion in *i*- and *u*-stems that applies before many of the same case endings that appear to trigger vowel syncope elsewhere. This latter process of vowel insertion is evidently driven by a need to provide syllable onsets while avoiding illicit consonant sequences or superheavy syllables. Compare the paradigms of the Ved. *u*-stems *krátu*- m. 'power' and *sūnú*- m. 'son':

| CASE | **krátu-** | **sūnú-** |
|---|---|---|
| nom.sg. | *krátuḥ* | *sūnúḥ* |
| acc.sg. | *krátum* | *sūnúm* |
| inst.sg. | *krátvā* | *sūnávā* |
| dat.sg. | *krátve* | *sūnáve* |
| abl./gen.sg. | *krátvaḥ* | *sūnóḥ* |
| nom.pl. | *krátavas* | *sūnávaḥ* |
| acc.pl. | *krátūn* | *sūnū́ṃs* |
| inst.pl. | *krátubhiḥ* | *sūnúbhiḥ* |
| loc.pl. | (*krátuṣu*) | *sūnúṣu* |

Note that, before consonantal endings, there is no difference in the form of the stem; before the vocalic endings (inst.sg. /-ā́/, dat.sg. /-é/, nom.pl. /-as/), a vowel *a* is inserted in the stem of *sūnú*-, creating *sūnáv*-. After the same fashion, zero grade in *r*- and *n*-stems might serve to preserve alignment between syllables and morphemes and to avoid (super)heavy syllables.[15]

In short, while some healthy correlation between inherently accented morphemes and zero grade ablaut is present in the Vedic lexicon, its precise status is uncertain. On account of its opacifying effects, it is empirically worrisome. Furthermore, the possibility of accounting

[15]The possibility of deriving ablaut patterns through constraints on optimal foot structure seems promising in the abstract, but a small-scale attempt in Keydana 2014 appears unsuccessful to me.

for the same effects through conditions on optimal syllabification means that it cannot alone be taken as a reliable indicator of a morpheme's accentual properties. At present, I will then eschew the use of abalut or apparent ablaut as a diagnostic for accentuation. More generally, the synchronic status and conditioning of ablaut processes in Sanskrit ought to be a topic of more systematic investigation.

### 7.1.5 Accentual Mobility

Examples (14)–(18) above demonstrate that nominal paradigms in Greek and Vedic may exhibit accentual mobility, depending upon the accentual properties of the root and suffix. In those examples, mobility falls out strictly from a combination of unaccented and accented morphemes, plus a default phonological ictus. However, accentual mobility in these languages is strikingly limited – indeed, in Greek, truly mobile nominal paradigms that can be analyzed and accounted for as with /pod-/ 'foot' in (18) are limited to monosyllabic root nouns.[16] The same is nearly true for Vedic as well, though at least two other forms, *pánthā-* 'path' and *púmaṃs-* 'man', which are (largely) disyllabic throughout their paradigms (cf. 7.2.2 below), also exhibit a mobile ictus.[17] Despite the logical possibility of an unaccented root combining with an unaccented derivational suffix, and thus forming an unaccented stem, when combined with an accented inflectional ending, the automatic result is **not** ictus movement between the root syllable and the inflectional ending. Why this logical possibility should be so rarely, if ever, realized, requires an answer. For instance, I will present reasons at 7.2.2 and 7.2.3 to believe that the Sanskrit derivational suffix /-vant-/ ('having X') is unaccented. In combination with an unaccented root such as /pad-/ 'foot' to form the adjective /pad-vant-/, the combination of the stem /pad-vant-/ with the accented inflectional ending gen.sg. /-ás/ gives a form *padvátas*, with ictus on the syllable of the derivational suffix, rather than the ending, $^{X}$*padvatás*.

Before pressing further, one must recognize that both languages exhibit two entirely distinct patterns of PSEUDO-MOBILITY of the ictus. Attic-Ionic pseudo-mobility results simply from the constraints imposed by the Law of Limitation and the optimal recessive ic-

[16] In fact, inflectional stems that become monosyllabic through historical processes of consonant lenition and vowel contraction in Greek take on mobile accentuation. For instance, Proto-Greek *[ówhos], gen.sg. *[ówhatos] 'ear' appears in Homer as οὖς [ô:s], gen.sg. οὔατος [ó:atos], with persistent high tone on the first syllable, though Attic attests the further contracted gen.sg. ὠτός [ɔ:tós], and accentual mobility. See further Rix 1976 [1992]: 148 and de Lamberterie 2009: 92 ff..

[17] Besides those two examples, I know of no other clear examples of ictus movement between the root syllable and an inflectional ending in Vedic in a stem that is polysyllabic in the strong cases. Feminine *ī*-stems of the *devī́*-type (nom.sg. *devī́*, gen.sg. *devyā́s* 'goddess') are often taken to continue an Indo-European pattern with ictus on the root in strong cases and ictus on the derivational suffix */-i̯eh$_2$-/ in the weak cases (cf. Meier-Brügger 2003: 285–7). A survey of feminine *ī*-stems in the RV with ictus on the initial syllable in the nom.sg., however, returns no items that show any movement of the ictus whatsoever.

Worth noting, though, is the peculiar ictus pattern found in some numerals (though these patterns are surely innovative, and the rarity of the oblique case forms suggests that they are productively built): nom./acc. *náva* 'nine', gen.pl. *navānā́m*, but inst.pl. *navábhiḥ*; similarly *páñca* 'five', gen.pl. both *pañcānā́m* and inst.pl. *pancábhiḥ*, loc.pl. *pañcásu*. Similar in this respect are neuters like *ákṣi* 'eye' and *ásthi* 'bone', which inflect like derived *n*-stems with pseudo-mobility (gen.sg. *akṣnás*, inst.pl. *akṣábhiḥ*) outside the nom./acc.sg.

tus: when a form within a paradigm changes in number of morae, the position of the ictus may shift accordingly. Thus we find nom.sg. ἄνθρωπος [ánt$^{h}$rɔ:pos], but acc.pl. ἀνθρώπους [ant$^{h}$rɔ́:po:s]. Attic-Ionic pseudo-mobility thus follows from the constraints on the placement of the phonological ictus. In Vedic, pseudo-mobility is a consequence of the change of syllabic /ṇ/ (realized as [a]), [u], [i], or [ṛ], which can bear the ictus, into non-syllabic [n], [w], [j], or [r], which cannot, in which case the surface position of the ictus shifts immediately rightward. This behavior is evident in an *n*-stem such as *ukṣán-* 'bull': acc.sg. /ukṣ-án-am/ →[ukṣánam], gen.sg. /ukṣ-án-ás/ → [ukṣnás] (zero grade applies to the derivational suffix, and /n/ fills an onset), inst.pl. /ukṣ-án-bhís/ → [ukṣábhis] (zero grade applies to the derivational suffix, and /n/ acts as a syllable nucleus). These types of accentual mobility, which are commonplace in the respective languages, do not pose a problem in themselves.

The Greek pseudo-mobility induced by the Law of Limitation, however, precisely means that lexical accents that potentially violate the Law of Limitation will be compelled to have pseudo-mobility. For instance, if one were to analyze the stem /ánt$^{h}$rɔ:po-/ as having a lexical accent as shown, the nom.sg. [ánt$^{h}$rɔ:pos] would allow for faithful realization of that accent, but the position of the ictus in acc.pl. [ant$^{h}$rɔ́:po:s] would be as close to the position of the lexical accent as permitted. The fact that the gen.pl. appears as [ant$^{h}$rɔ́:p:ɔ:n] rather than $^{X}$[ant$^{h}$rɔ:p:ɔ̂:n] could then potentially be explained as faithfulness to the accent of the stem /ánt$^{h}$rɔ:po-/ over the accent of the inflection /-ɔ̂:n/. Such an explanation, however, would be tantamount to assuming that *all* polysyllabic nominal stems in Greek are lexically accented, the consequence of which is that a possible lexical contrast in the language (i.e., between inherently accented and unaccented polysyllabic stems) would be surprisingly absent. Furthermore, forms that Probert (2006b) explains as exhibiting recessive accentuation via historical "deaccentuation" (e.g., */akró-/ 'highest point' > /akro-/) would not be truly "deaccented" in the sense that they would have no lexical accent, but just a different one.

Steriade (2014) (as a "stop-gap solution") indeed suggests a similar possibility, but which does not constrain the inputs: a constraint Polysyllabic requires the high toned part of the ictus to fall on the stem, if the stem is polysyllabic. This constraint is descriptively adequate, but is *ad hoc*, and it seems peculiar to me to operate with a constraint that effectively counts the number of syllables in a stem (or at least can distinguish between 1 and greater than 1).[18] A less disquieting solution might be the use of Output-Output Correspondence between the surface accentuation of the nom.sg. and other members of the paradigm (i.e., given the accentuation of [ánt$^{h}$rɔ:pos], other parts of the paradigm strive to maintain the same accentuation, insofar as permitted by the Law of Limitation). A strict Output-Output Correspondence, would, however, undergenerate with respect to the mobile monsosyllabic nouns (i.e., it would predict gen.pl. $^{X}$[pódɔ:n] on account of nom.sg. [pó:s] or acc.sg. [póda]). Let us define the possible Output-Output constraint as in (26).

(26) Ident-OO(Ictus, nom.sg.): Assign one violation for each syllable by which the position of the ictus differs from ictus position in the nom.sg. of the same lexeme.[19]

[18] Kiparsky (Forthcoming), for instance, criticizes a Slavic accentual sound law proposed by Jasanoff 2008 precisely for relying on a syllable counting condition.

[19] Practically speaking, this Output-Output constraint could refer to the position of the ictus in any strong

The effect of this constraint is probably identical to Steriade's POLYSYLLABIC, but saves the generalization that the position of the ictus in the nom.sg. may be relevant.[20] An identical constraint in Vedic would limit accentual mobility in just the same way, and predict that, in case a stem vacilates in length between one and two syllables, mobility might be possible; I referred to two such cases above. Perhaps the only instance of ictus mobility in Vedic that spans a distance greater than one syllable across inflected forms occurs in reduplicated presents, where the ictus may fall either on the initial syllable (1.sg. *bíbharmi* 'I bear') or the inflectional ending (1.pl. *bibhṛmási* 'we bear').

Given the core element of inherently accented inflectional morphemes as drivers of accentual mobility, the rarity of genuinely mobile nouns in Greek and Vedic is surprising. A constraint along the lines of Steriade's POLYSYLLABIC or IDENT-OO(Ictus, nom.sg.) proposed here could be sufficient to limit mobility essentially to monosyllabic stems. A more thorough survey of the Greek and Vedic lexicons, to draw out other telling instances of accentual mobility, would be worthwhile.

### 7.1.6 Logical Possibilities for the Analyst

Table 7.1 above should already give a sense of the multifarious analytical possibilities that present themselves in attempting to decide why a give lexeme exhibits its particular accentuation. To take a relatively simple example, a neuter *s*-stem such as Skt. *mánas-* 'thought', with ictus on the initial syllable, could show that ictus because: the root /mán-/ is accented (/mán-as-/); the derivational suffix /-ˊas-/ is pre-accenting (/man-ˊas-/); or both root and derivational suffix are unaccented, and the ictus reflects Sanskrit's phonological default (/man-as-/). From among those three possibilities, there is the further possibility of seeing a composite stem, either accented /mánas-/ or unaccented /manas-/, rather than a morphologically analyzed stem. Likewise, a feminine *ti*-stem such as Skt. *matí-*, with ictus on the second syllable, which etymologically contains the same root /man-/, could show that ictus because: the derivational suffix /-tí-/ is accented and dominant, and thus receives the ictus regardless of whatever accentual property /man-/ may have (/man-tí-/ or /mán-tí-/), and conditions zero-grade ablaut of the root; or reflects a composite lexicalized stem /matí-/. For polysyllabic stems that show the ictus on the final syllable of the stem (such as *padvánt-* 'footed, having feet' cited above), Kiparsky 2010: 144 opens yet another option:

(27) OXYTONE RULE
$\sigma \rightarrow \acute{\sigma}$ [...$\sigma$ __]$_{\text{Stem}}$ - Inflection
An accent is assigned to the edge of all polysyllabic inflected stems.

Thus *padvánt-* could reflect /pad-vant-/, which would receive an accent by the Oxytone Rule, giving /padvánt-/, and ultimately show a second syllable ictus in *padvánt-*. The Oxytone Rule, as formulated here, would be redundant in case the rightmost vocalic element in a

case form. Below, I will use the form of the acc.sg. as the paradigmatic base, but one could easily choose the nom.sg. or nom.pl. without affecting the predictions of the grammar.

[20] Compare Vendryes 1904 [1945]: 206: "l'accent conserve dans la flexion la nature et la place qu'il possède au nominatif singulier".

stem already possessed a lexical accent. By chapter's end, the objective is to make the distinctions and motivations behind the analytical choices more well-defined and well-motivated. In particular, the ability to systematically distinguish morphologically parsable forms from morphological simplexes should offer insight into the accentuation of specific lexemes.

## 7.2 Accentual Dominance and Headedness

As a first step towards better understanding why Greek and Vedic accentuation is constrained in certain ways, while being able to motivate systematic interactions between accentuation and morphology, I will borrow heavily from the work of Revithiadou 1999. Revithiadou's research demonstrates that, in numerous languages, ranging from Modern Greek and Russian to Salishan and Yupik languages, the *head* of a morphologically complex word plays a central role in determining how prosodic prominence is realized. The relevance of headedness to the study of accentual changes and seeming accentual irregularities will quickly become evident: where a lexeme that was etymologically headed by a given morpheme loses that head, different accentuation may result. Consideration of productivity and other factors treated in Chapter 3 that help to distinguish between genuinely complex and merely etymologically complex lexemes can then further clarify exactly when an apparent morpheme truly heads a lexeme, or is a mere historical residue.

The requisite now is a clear definition of the notion "head of the word", so that the set of elements for which an effect of headedness might be expected can be defined. Most simply, the head of a morphologically complex word is the component of the word that specifies certain properties of the word as a whole that compose the subcategorization frame of that word. Those subcategorizational properties are crucial, because they determine how a word may interact syntactically with other words or combine with other morphemes. The relevant properties are then the syntactic category of the word (e.g., noun, verb, etc.) and gender/class, which may compel agreement phenomena or require the idiosyncratic selection of certain inflectional morphemes. Thus, in the broadest sense, morphemes typically labeled as derivational will often be morphological heads, while inflectional morphemes will never be morphological heads. Whether roots alone are to be considered heads on par with derivational morphemes is less clear. The general association between derivational morphology and headedness, and inflection and absence thereof, is essentially the same definition offered in Zwicky 1985 and Zwicky 1993.

The claim of Steriade 1988a mentioned at 7.1.3 above, that all derivational suffixes in Greek are dominant, could now find a sensible and constrained interpretation, without needing to introduce a further lexical diacritic [+dominant], or necessarily relying on a morphophonological cycle to ensure the application of dominance effects: accentual dominance would merely be the preferential expression of a head's accentual properties to the exclusion of other underlying accents. As such, the appropriate ranking of Input-Output faithfulness constraints that make reference to headedness (in effect, that can partially "see" inside the structure of morphologically complex words output by the lexicon) can ensure accentual dominance effects, all the while operating within a single phonological stratum. For conve-

nience, however, I will continue to use the term 'dominance effect' to refer to the seeming "erasure" of other lexical accents by some morpheme.

### 7.2.1 HeadFaith and HeadStress (with reference to Tōkyō Japanese)

Revithiadou (1999: 26–31) describes two major ways in which morphological heads may receive prosodic prominence: either the lexical accents of a head may be preferentially realized over other lexical accents, or prosodic prominence may be assigned directly to the head, to the exclusion of other competing prosodic possibilities. These two ways in which surface prosodic prominence may come to be associated with a morphological head is expressed through two constraints, HeadFaith and HeadStress.

(28) HeadFaith
A lexical accent sponsored by a morphological head in the input has a correspondent in the output (HeadMax(Accent)).
A lexical accent hosted by a morphological head in the output has a correspondent in the input (HeadDep(Accent)).

(29) HeadStress
Morphological heads are stressed (= receive the ictus).

In case the specific HeadFaith constraint outranks general faithfulness (Faith) to lexical accents (i.e., Max-IO(Accent) and Dep-IO(Accent)), dominance effects can be obtained; Revithiadou (1999: 29) says that Modern Greek and Russian are languages in which the ranking HeadFaith ≫ Faith ≫ HeadStress obtains. The consequence is that the lexical accents of heads in those languages win out over the lexical accents of non-heads, but heads are not automatically assigned the ictus. In Ancient Greek, the behavior of the suffix [-ísko] shown in (24) above can follow from HeadFaith: in the derivation /as-tér-ísko-s/ → [aster-ískos] 'little star', the accent of /-ísko-/ takes precedence over the accent of /-tér-/, because /-ísko-/ is the head of the entire lexeme. By implication, if a form like [alέːtʰeia] 'truth' indeed reflects a morphologically parsed /alɛːtʰ-éː-ia/, then the proper analysis of the affix /-ia-/ is as an unaccentable suffix, which forces the ictus outside of its domain. Were /-ia-/ merely unaccented, one might expect to find the accent of /-éː-/ surface faithfully, giving *[alɛːtʰéia].

A straightforward case of a HeadStress system presents itself in the accentuation of Thompson Salish compounds (see discussion in Revithiadou 1999: 263 ff.; data are from Thompson and Thompson 1996). The semantic head of the compound invariably takes the word ictus, regardless of any underlying accentual properties. The head in these cases is always the rightmost member of the compound.

(30) Thompson Salish Root + Lexical Suffix Compounds
a. /ƛ̓'ixʷel+xən/ → ƛ̓'ixʷeɬ-xə́n 'different shoes' (Thompson and Thompson 1996: 182)
$\sqrt{\text{different}}$+shoe

b. /s-weʔwít+xən/ → s-wewit-xán ‘lower part of hind foot or leg’ (Thompson and Thompson 1996: 372)
NOM-$\sqrt{\text{behind}}$+foot

c. /p’uƛ’+qín/ → p’uƛ’-qín ‘foggy mountain top’ (Thompson and Thompson 1996: 261)
$\sqrt{\text{misty}}$+top

d. /sip’éc’+qín/ → sip’ec’-qín ‘scalp (noun)’ (Thompson and Thompson 1996: 327)
$\sqrt{\text{skin}}$+head

e. /kawpúy+esk’iʔ/ → kawpuyh-ésk’iʔ ‘cowboy song’ (Thompson and Thompson 1996: 82)
$\sqrt{\text{cowboy}}$+song

The processes of ictus assignment Tōkyō Japanese, meanwhile, presents an interesting case. In general, derivational processes appear to reflect the ranking HEADFAITH ≫ FAITH. The accentuation of noun-noun compounds, however, appears to constitute a subgrammar in which HEADSTRESS plays an important role. Observations concerning Japanese set the stage for the analysis of Sanskrit prosody as a system in which HEADSTRESS is relevant and necessary, but of Ancient Greek prosody as a system in which HEADFAITH is the deciding constraint.

Data in examples (31)–(33) are from Tsujimura 2014: 99–102. First, consider the suffix /-te/, which forms a verbal noun (traditionally labeled “gerund”) used in a great variety of constructions including in-progress action and change of state.[21] When combined with a verbal root that contains a lexical accent, that accent surfaces; when combined with an unaccented verbal root, the default tonal pattern (cf. example (21) above) arises. /-te/ is a morphological head, but because it is unspecified for any prosodic properties, general faithfulness to other underlying accents can be satisfied. Since we do not find an ictus (high-to-low tone downstep) assigned to the head [-te] in surface forms, we should conclude that HEADSTRESS is low-ranked.

(31) Unaccented suffix /-te/ (gerund)[22]
a. /yóm-te/ → yónde ‘read’
b. /omów-te/ → omótte ‘think’
c. /ugók-te/ → ugóite ‘move’
d. /káer-te/ → káette ‘return’
e. /tábe-te/ → tábete ‘eat’
f. /mí-te/ → míte ‘see’
g. /áe-te/ → áete ‘mix’
h. /kazóe-te/ → kazóete ‘count’

[21] See Kaiser et al. 2001 [2013]: 216–26 on uses of /-te/.

[22] Acute here indicates the ictus = “downstep”, i.e., the syllable after which all tones are low. The same lexical glosses given here apply to the following two example sets.

i. /sí-te/ → síte ‘force’
j. /kik-te/ → kīte ‘listen’
k. /ake-te/ → akete ‘open’
l. /sawar-te/ → sawatte ‘touch’
m. /ire-te/ → irete ‘put in’

In contrast, the derivational suffix /-(y)ó/, usually described as “hortative” (e.g., *ikimasyó* ‘let’s go!’; cf. Kaiser et al. 2001 [2013]: 229–37), always bears the ictus precisely on that derivational suffix. Head faithfulness is fully satisfied at the expense of lexical accents borne by roots.

(32) Accented derivational suffix /-(y)ó/ (“tentative” or “hortative”)
a. /yóm-yó/ → yomó
b. /omów-yó/ → omoó
c. /ugók-yó/ → ugokó
d. /káer-yó/ → kaeró
e. /tábe-yó/ → tabeyó
f. /mí-yó/ → miyó
g. /áe-yó/ → aeyó
h. /kazóe-yó/ → kazoeyó
i. /sí-yó/ → siiyó
j. /kik-yó/ → kikó
k. /ake-yó/ → akéyó
l. /sawar-yó/ → sawaró
m. /ire-yó/ → iréyó

The behavior of unaccented /-te/ and accented /-(y)ó/ may in turn be contrasted explictly with the clause-final clitic /+tára/, which forms conditional clauses ‘when, if’.[23] /+tára/ inevitably forms a prosodic word with a a verb, e.g., /yóm+tára/ → *yóndara* ‘if he reads’. Since /+tára/ is not the morphological head of the form *yóndara*, its lexical accent surfaces only when it falls within a prosodic word that is otherwise unaccented, e.g., /sawar+tára/ → sawattára ‘if he touches’.[24]

(33) Pre-accenting clausal clitic /-tára/ (conditional)
a. /yóm+tára/ → yóndara

[23]This affix is traditionally described as a suffix *-tara* that attaches to a verbal root, but since it takes scope over the whole clause, it is most appropriately described as a clausal clitic. On uses of *-tara*, see Kaiser et al. 2001 [2013]: 575–7.

[24]Thus, *pace* Alderete (2001: 105), it is not the case that the lexical accents of roots are privileged over the accent of suffixes in Japanese; the competition between lexical accents in /yóm+tára/ is decided by the phonology, which seems to prefer for the ictus to be towards the left edge of the prosodic word.

b. /omów+tára/ → omóttara
c. /ugók+tára/ → ugóitara
d. /káer+tára/ → káettara
e. /tábe+tára/ → tábetara
f. /mí+tára/ → mítara
g. /áe+tára/ → áetara
h. /kazóe+tára/ → kazóetara
i. /sī́+tára/ → sī́tara
j. /kik+tára/ → kītára
k. /ake+tára/ → aketára
l. /sawar+tára/ → sawattára
m. /ire+tára/ → iretára

Tōkyō Japanese endocentric Noun + Noun compounds, meanwhile, exhibit a strong tendency to maintain or place an ictus on the semantic head of the compound, insofar as permitted by other phonological constraints. The semantic head, in these cases, is always the righthand member of the compound. Data here derives from Kubozono 1995, Alderete 2001: 106, and Tsujimura 2014: 85 ff.). The generalizations concerning ictus placement appear to be as follows:

1. final syllables are extrametrical, and therefore cannot sponsor an ictus;
2. the accent of the head (the second member) persists if inherently accented on a non-final syllable;
3. if the head is inherently accented on the final syllable, or has no inherent accent, and has at least two morae preceding the last syllable, assign a default accent to the first syllable of the head;
4. if the head does not contain at least two morae before the final syllable, assign a default accent to the final syllable of the first member, thus marking the compound boundary.[25]

(34) The underlying accent of the head is maintained:
a. /yamá+hototógisu/ → yama-hototógisu 'mountain quail'
b. /síritu+syōgákkō/ → siritu-syōgákkō 'private elementary school
c. /ō+kámakiri/ → ō-kámakiri 'big mantis'
d. /nise+karakása/ → nise-karakása 'paper umbrella'
e. /dorobō+néko/ → dorobō-néko 'thief cat = a pet cat that steals food from other houses'

[25] A significant residue of apparent exceptions remains, but these principles seem to capture the productive patterns.

f. /hukuró+néko/ → hukuro-néko ‘bag cat = a kind of marsupial (genus Dasyurus)’

g. /ko+néko/ → ko-néko ‘child cat = kitten’

(35) Default accent assigned to the first syllable of the head:

a. /kyó+yasai/ → kyō-yásai ‘vegetable from Kyōto’

b. /minami+amerika/ → minami-ámerika ‘South America’

c. /áisu+kōhí/ → aisu-kóhī ‘iced coffee’

d. /té+kagamí/ → te-kágami ‘hand mirror’

(36) Default accent assigned to last syllable of the non-head, i.e., as close to the first syllable of the head as the phonology permits.

a. /kuwagáta+musi/ → kuwagatá-musi ‘stag bug = stag beetle’

b. /kábuto+musi/ → kabutó-musi ‘helmet bug = beetle’

c. /ákita+inú/ → akitá-inu ‘Akita dog’

d. /kensetu+syō/ → kensetú-syō ‘Ministry of Construction’

(37) No default downstep accent assigned, because there is no phonologically licit recipient of an ictus in trimoraic forms (a high tone is dispreferred on an initial syllable, and the ictus is kept away from the last two morae):

a. /huyú+kí/ → huyu-ki ‘winter tree’

b. /goma+sú/ → goma-su ‘sesame vinegar = sauce with sesame seeds’

c. /ko-umá/ → ko-uma ‘child horse = foal’

The most crucial distinction between the compound forms and simply derived lexemes in Tōkyō Japanese is that, while derivational suffixes do not automatically attract the ictus (cf. /-te/ in (31) above), heads of compounds do. Outside compounds, it appears that the general ranking HeadFaith ≫ Faith ≫ HeadStress is active; in compounds, the addition of a higher ranked constraint of greater specificity, $\text{HeadStress}_{\text{Compound}}$, ranked over general Faith, will provide for the patterns illustrated in examples (34)–(37). In the following sections, I will show that HeadStress plays an active role in Sanskrit, and that in fact HeadFaith is superfluous in that system, the effects thereof being derivable from the ganging of general Faith and HeadStress violations.

### 7.2.2 Tracing the Path of the Oxytone Rule

Although the default phonology ictus of Vedic Sanskrit appears to be driven by Align-L(Ictus, PrWd) (cf. examples (14)–(18) above), polysyllabic words very commonly instead exhibit a persistent ictus (or ictus and pseudo-mobility) on the syllable at the right edge of the stem. Precisely to capture this generalization, Kiparsky (2010: 144) formulates a rule that assigns an accent to the edge of all polysyllabic stems, which we encountered first at (27) above.

(38) Oxytone Rule (OR)
$\sigma \rightarrow \acute{\sigma}$ $[\ldots\sigma$ __$]_{\text{Stem}}$ - Inflection
An accent is assigned to the edge of all polysyllabic inflected stems.

The ultimate effect of the OR is to create stems with penultimate ictus, just in case no syllable further to the left has an underlying accent; Kiparsky orders the OR at the end of the stem-level phonology, adding an accent, before the post-lexical BAP selects from among accents to realize as the ictus. For example, Kiparsky (2010) considers the agent noun suffix /-tar-/ to be unaccented, and thus derives the accentuation of *pitáram* acc.sg. 'father' as follows: /pi-tar-/ → /pi-tár-/ (OR) → /pi-tár-am/ → [pitáram] (BAP). Accentually, when in combination with an unaccented root, polysyllabic stems with unaccented derivational suffixes thus become indistinguishable from those with accented derivational suffixes.

This rule, however, is subject to some troubling counterexamples. A number of classes of derived nouns and adjectives that result in polysyllabic stems (nouns in *-ti-*, adjectives in *-mant-*, *-vant-*, and *-ra-*) all contain members that show initial syllable ictus when attached to roots that cannot be underlyingly accented. For instance, the RV attests *śrúti-* 'boon' three times; the root $\sqrt{śru}$ 'listen', however, is most likely not inherently accented, because the Vedic also attests *śrutí-*, which would be categorically predicted by the Oxytone Rule.[26] The existence of parallel forms that differ in accentuation indicates that whatever process is responsible for systematic ictus at the right edge of many stems, it cannot depend simply on polysyllabicity of the stem.

A similar problem afflicts the accentuation of two lexemes with unique inflection in the language: *púmaṃs-* 'man' and *pánthā-* 'path'; their paradigms, as attested in the RV, are given in (39).[27] These two lexemes have polysyllabic stems in the strong cases, yet show ictus mobility, between the first syllable of the stem and the inflectional ending.

(39) *pumāṃs-* and *panthā-*

| **Case** | /pumās-/ | | /panthā-/ | |
|---|---|---|---|---|
| | Sg. | Pl. | Sg. | Pl. |
| Nom. | púmān | púmāṃsas | pánthās | pánthās, pánthānas |
| Acc. | púmāṃsam | puṃsás | pánthām | pathás |
| Inst. | – | – | pathā́ | pathíbhis |
| Gen. | puṃsás | puṃsā́m | pathás | pathā́m |
| Loc. | puṃsí | puṃsú | pathí | pathíṣu |

Since /pumaṃs-/ and /panthā-/ have polysyllabic stems, an accent should be assigned to the second syllable, but at the same time, the mobility of the ictus between the first syllable and the inflectional endings, as in gen.sg. *puṃsás* and *pathás*, excludes the possibility of an

[26] Note also that all nouns derived with the suffix *-ti-* to which a further derivative in *-mant-* is attested in the RV always show the ictus on the suffix *-mánt-* (e.g., *svatimánt-*). The implication is that the suffix /-ti-/ is unaccented. See further the following subsection 7.2.3 on *-mant-*.

[27] The forms inst.pl. *pathíbhis* and loc.pl. *pathíṣu* pose their own special problems. The *i* seen in these forms continues a PIIr. anaptyctic vowel; the immediate surface preform of *pathíṣu* could be PIIr. *[pət$^{h}$ɨšu]. In Sanskrit, the resulting *i* was evidently reanalyzed as the derivational suffix /-i-/, and indeed other *i*-stem inflecting forms, e.g., nom.pl. *patháyas*, are attested later. The reanalysis as the derivational suffix *-i-* would also account for the synchronic accentuation of the inst.pl. and loc.pl.: the ictus is assigned to the suffix /-i-/ to satisfy HeadStress.

inherent accent on the first syllable of the stem. If the OR functioned as Kiparsky formulates it, the expected accentuation of these two lexemes would be as in (40) below.

(40) *pumāṃs-* and *panthā-* with accentuation by OXYTONE RULE

| **Case** | /pumās-/ | | /panthā-/ | |
|---|---|---|---|---|
| | Sg. | Pl. | Sg. | Pl. |
| Nom. | [X]**pumā́n** | [X]**pumā́ṃsas** | [X]**panthā́s** | [X]**panthā́s, panthā́nas** |
| Acc. | [X]**pumā́ṃsam** | puṃsás | [X]**panthā́m** | pathás |
| Inst. | – | – | pathā́ | pathíbhis |
| Gen. | puṃsás | puṃsā́m | pathás | pathā́m |
| Loc. | puṃsí | puṃsú | pathí | pathíṣu |

Two crucial facts about the morphological composition of *panthā-* and *pumaṃs-* render them distinct from most other polysyllabic nouns in Vedic, however. First, no other lexeme in the language appears to contain material that could be neatly equated to a derivational suffix in either word. *panthā-* is the only masculine gender noun in the language that terminates in an ablauting stem *-ā*. Furthermore, the etymological root on which each word is based is absent from any other nominal or verbal form.[28] Thus neither stem would appear to contain a clearly segmentable root and derivational suffix. These two words must reflect simplex (root-like) stems, which cannot be synchronically decomposed any further: /panthā-/ and /pumaṃs-/.[29]

The ready conclusion that *panthā-* and *pumaṃs-* are morphologically simplex offers the opportunity for a reinterpretation of Kiparsky's Oxytone Rule in the light of Revithiadou's work: most apparent instantiations of the Oxytone Rule could in fact reflect the effects of a HEADSTRESS constraint. That is, a large number of Vedic lexemes that exhibit a persistent ictus at the right edge of the stem show that ictus because the right edge of the stem regularly corresponds to the outermost derivational suffix. Allowing for the decomposition of /pi-tar-am/ as above, the correct position of the ictus in the acc.sg. *pitáram* can then be derived in a single phonological stratum by weighting or ranking HEADSTRESS above ALIGN-L(Ictus, PrWd).[30] Conversely, for a morphologically simplex stem combined with an unaccented inflectional ending, such as /pumaṃs-am/ or /pad-am/, ALIGN-L(Ictus, PrWrd) alone decides the position of the ictus. To obtain paradigmatic mobility, and produce forms

[28]The root of *panthā-* is a PIE */pont-/, reflected in Gk. πόντος [póntos] 'sea', Lat. *pons, pontis* 'bridge', Armenian *hown* [hun] 'bridge'; the derivational suffix would seem to be a PIE */-oh$_2$-/. *pumaṃs-* may be a compound or univerbation in origin of */pu-/ 'young' (cf. Ved. *putrá-* 'boy, son') and */mas-/ 'man' (cf. Lat. *mās* 'man'). See Mayrhofer 1986–2001: s.v.v. *pánthā-* and *púmaṃs-* for further general etymological information, and Garnier 2010 for a detailed discussion and another proposal.

[29]To be clear, the preceding discussion may be somewhat anachronistic. The unusual double application of zero grade, to derive the stem *path-* found in, e.g., the gen.sg. *pathás*, from /panthā-ás/, is difficult, perhaps impossible, to motivate synchronically in Vedic.

[30]Giving a parsed stem derivation /pi-tar-am/ is somewhat infelicitous, because Vedic kinship terms have some unique patterns of inflection that indicate that they should be treated separately from the agent noun suffixes /-tár-/ and /-ˊtar-/ (/-ˊtar-/ here generates root-accented forms like Ved. *dā́tar-*). In fact, that *pitár-* has an underlying accent is shown by the form *pitṛ́mant-* (AVŚ 2× dat.sg. *pitṛ́mate*) ← /pitár-mant-/.

like gen.sg. *puṃsás*, the question is merely the correct weighting of ALIGN-L(Ictus, PrWrd), MAX-IO(Accent), and IDENT-OO(Ictus, nom.sg.). The overall behavior of the Vedic system becomes clear with the examination of the productive derivational suffixes /-mant-/ and /-vant-/.

### 7.2.3 The Suffixes /-mant-/ and /-vant-/

The Vedic suffixes *-mant* and *-vant* (the latter being etymologically comparable to Gk. nom.sg. -(ϝ)ειϛ [-(w)e:s], gen.sg. -(ϝ)εντος [-(w)entos] and Hitt. *-want-*) are very high type frequency suffixes, which attest numerous hapax legomena in the RV, that form adjectives from nouns with the meaning 'possessing X'. Kiparsky (Forthcoming) explicitly identifies these two suffixes as being accented and recessive; hence, in combination with a base, an ictus would be predicted to surface on *-mánt-* or *-vánt-* only if the base contains no lexical accent of its own. This analysis is logically possible: we find *padvánt-* 'footed, having feet', thus /pad-vánt-/, and *gómant-* 'having cows', thus implying /gó-mánt-/, in line with the accentual behavior of the root nouns /pad-/ and /gáv-/. From the theoretical point of view espoused here, however, the existence of a category-changing affix that does not exhibit dominance effects in Vedic is problematic: if the affix meets the definition for headhood, then, if it is specified for some prosodic property, it ought to impose that property consistently. Given the probably high productivity of these two suffixes, there is also no reason to believe that we find substiantial variation of the ictus position due to lack of parsability in the forms that instantiate it (cf. discussion in 7.3 and 7.4 below).

The reformulation of the Oxytone Rule as above, and its interpretation as satisfying HEAD-STRESS, however, allow for a different analysis: /-mant-/ and /-vant-/ are rather unaccented suffixes, that often receive the ictus to satisfy HEADSTRESS when other accents further to their left are not present. We can now write a small accentual grammar of Vedic that accounts for the following behaviors:

- accentual mobility in nouns such as acc.sg. *pā́dam*, gen.sg. *padás* or acc.sg. *púmāṃsam*, gen.sg. *puṃsás*. Note, in particular, that all mobile nouns in Vedic are disyllabic in the oblique cases.
- persistent ictus on the root in forms such as acc.sg. *gā́vam*, gen.sg. *gávas* or acc.sg. *gómantam*, gen.sg. *gómatas*.
- persistent ictus on the derivational suffix in forms such as acc.sg. *padvántam*, gen.sg. *padvátas*.
- shift of ictus to a derivational suffix in forms such as PPP *taṣṭá-*, where the ictus remains on the root in inflectional forms (3.pl.pres. *tákṣ-ati* 'they fashion' vs. 3.pl.pres. *s-ánti* 'they are').

The analysis here is formulated in a simple version of Harmonic Grammar (Smolensky and Legendre 2006). This analytical choice is crucial, because it permits for the weights of

constraints to "gang", and thereby eliminate candidates that would win under the presumption of strict ranking.[31] The weights given for the constraints used here were determined by solving a system of linear inequalities over the set of candidates and their violation profiles, as discussed in Potts et al. 2010, using an implementation in version 5.4 of Praat (Boersma and Weenink 1992–2014).[32] Throughout, I assume undominated CULMINATIVITY (cf. example (12) above) to exclude candidates with no or multiple word accents; similarly, I leave out DEP-IO(Accent) because it never plays a crucial role in determining a winner. Numbers in the tableaux indicate the weight of a violation, and multiple violations are indicated as the weight of the constraint × number of violations. The summed "harmony score" of each candidate appears at the left; the candidate with the lowest harmony score is selected as the winner. Ultimately, a working analysis is possible with just four constraints:

(41) Constraints

a. MAX-IO(Accent): an accent in the input must have a correspondent in the output. **Weight**: 4

b. HEADSTRESS$_{\text{Affix}}$: the affixal head of a morphologically complex word must receive the ictus. **Weight**: 4

c. ALIGN-L(Ictus, PrWd): the syllable with the ictus must be aligned with the left edge of the prosodic word; each syllable intervening between the syllable with the ictus and the leftmost syllable of the prosodic word counts as a violation. **Weight**: 1

d. IDENT-OO(Ictus, acc.sg.): the syllable with the ictus must be the same as the syllable with the ictus in the acc.sg.; each syllable intervening between the syllable with the ictus and the syllable with the ictus in the acc.sg. counts as a violation.[33] **Weight**: 1.5

First, I will assign a weight of 1 to ALIGN-L(Ictus, PrWd). Thus in cases where no underlying accents are present, the ictus appears at the left edge of the prosodic word:

[31] In fact, the hand ranking MAX-IO(Accent), HEADSTRESS$_{\text{Affix}}$ ≫ IDENT-OO(Ictus, acc.sg.) ≫ALIGN-L(Ictus, PrWd) will derive the correct winners for all examples considered here, provided that one allows for the pooling of violation marks within ranked strata (a somewhat dubious theoretical choice). However, I choose instead to use Harmonic Grammar because the standard learning algorithm for constraint ranking in parallel Optimality Theory, Constrain Demotion (see Tesar and Smolensky 2000) cannot discover this feasible ranking (because it does not admit of the possibility of pooled violation marks). Constraint Demotion fails to discover any feasible ranking because all four constraints are sometimes "loser-preferring", and the algorithm is therefore unable to form a top-ranked stratum of constraints. Since I believe that constraint grammars should be computationally learnable, the failure of Constraint Demotion on this data set leads me to use Harmonic Grammar instead.

[32] Specifically, I used the `LinearOT` decision strategy to `Find positive weights`. The number of forms that I used in solving for the constraint weights used here includes several more types than discussed here, so the model has in fact been tested on data with more empirical coverage, though most reproduce the same violation patterns as the examples below.

[33] I choose the acc.sg. as the reference point for this output-output constraint, but reference to any "strong" case form (nom.sg., nom.pl.) would have the same effect. I discussed the common idea of the nom.sg. as the paradigmatic base above, but since the acc.sg. forms are more transparent, and the nom.sg. and acc.sg. always show an identical ictus position in all paradigms in Greek and Sanskrit, I employ the acc.sg. here.

(42) acc.sg. *pā́dam* 'foot' ← /pād-am/

| acc.sg. pād-am | | | ALIGN-L(Ictus, PrWd) |
|---|---|---|---|
| a. | ☞ 0 | pā́dam | |
| b. | 1 | pādám | 1 |

By assigning a higher weight to MAX-IO(Accent), 4, we guarantee that the underlying accent is realized faithfully, if the resulting ictus is not too far from the left edge. Candidate b. wins in this case:

(43) gen.sg. *padás* 'foot' ← /pad-ás/

| pad-ás (BASE: *pā́dam*) | | | MAX-IO(Accent) | ALIGN-L(Ictus, PrWd) |
|---|---|---|---|---|
| a. | 4 | pádas | 4 | |
| b. | ☞ 1 | padás | | 1 |

However, in case the root morpheme at the left edge is itself accented, the "ganging" of a MAX-IO(Accent) and an ALIGN-L(Ictus, PrWd) violation will ensure realization of the ictus at the left edge. Candidate a. then wins in this case:

(44) gen.sg. *gávas* 'cow' ← /gáv-ás/

| gáv-ás (BASE: *gā́vam*) | | | MAX-IO(Accent) | ALIGN-L(Ictus, PrWd) |
|---|---|---|---|---|
| a. | ☞ 4 | gávas | 4 | |
| b. | 5 | gavás | 4 | 1 |

When an unaccented derivational suffix enters the fray, $\text{HEADSTRESS}_{\text{Affix}}$, with its weight of 4, it will decide the position of the ictus when no accents are underlyingly present.[34]

[34] At the weight of 1, three violations of ALIGN-L(Ictus, PrWd) will produce a harmony score of 3, but a violation of $\text{HEADSTRESS}_{\text{Affix}}$ has a weight of 4, and thus a form with up to three syllables preceding the suffix /-mant-/ or /-vant-/ will receive the ictus on that derivational suffix. In the RV, there are no forms with the ictus on *-mánt-* or *-vánt-* preceded by more than three syllables. It is an interesting question, whether any forms with four or more syllables preceding an unaccented derivational suffix ever receive the ictus on that suffix; an answer to that question would help to determine whether some extreme number of violations of ALIGN-L(Ictus, PrWd) might ever outweigh $\text{HEADSTRESS}_{\text{Affix}}$. If not, then the weights of $\text{HEADSTRESS}_{\text{Affix}}$ and MAX-IO(Accent) must have some weight greater than the greatest number of ALIGN-L(Ictus, PrWd) violations ever incurred in a winner.

(45) acc.sg. *padvántam* 'footed' ← /pad-vant-am/

| pad-vant-am | | | HeadStress$_{Affix}$ | Align-L(Ictus, PrWd) |
|---|---|---|---|---|
| a. | 4 | pádvantam | 4 | |
| b. ☞ | 1 | padvántam | | 1 |
| c. | 6 | padvantám | 4 | 1 × 2 |

In case of an inherently accented root, such as /gáv-/ 'cow', violations of Max-IO(Accent) and Align-L(Ictus, PrWd) gang to exclude the candidate with ictus on the derivational suffix. The same effects seen up to this point could be obtained by the strict ranking Max-IO(Accent) ≫ HeadStress$_{Affix}$ ≫ Align-L(Ictus, PrWd), which would predict that underlying accents are always preferred to the insertion of accent to satisfy HeadStress$_{Affix}$.

(46) acc.sg. *gómantam* 'having cows' ← /gáv-mant-am/

| gáv-mant-am | | | Max-IO(Accent) | HeadStress$_{Affix}$ | Align-L(Ictus, PrWd) |
|---|---|---|---|---|---|
| a. ☞ | 4 | gómantam | | 4 | |
| b. | 5 | gomántam | 4 | | 1 |
| c. | 10 | gomantám | 4 | 4 | 1 × 2 |

That the strict ranking Max-IO(Accent) ≫ HeadStress$_{Affix}$ is inadequate is shown by the gen.sg. *padvátas*; that ranking would predict a winner $^{X}$*padvatás*, with ictus on an underlyingly accented inflectional suffix. Instead, even when an accented inflectional suffix is added, the ictus remains stable on the derivational suffix, despite the violation of Max-IO(Accent), on account of the additional violations of Align-L(Ictus, PrWd) and Ident-OO(Ictus, acc.sg.). [35]

(47) gen.sg. *padvátas* 'footed' ← /pad-vant-ás/

[35] Since both candidates b. and c. violate either HeadStress$_{Affix}$ or Max-IO(Accent), the same effects could be obtained by ranking those two constraints in the same stratum and allowing for the pooling of violation marks within a stratum. A tableau with this ranking, HeadStress$_{Affix}$, Max-IO(Accent) ≫ Ident-OO(Ictus, acc.sg.), Align-L(Ictus, PrWd) then picks the correct winner:

| pad-vant-ás (BASE: *padvántam*) | | | MAX-IO(Accent) | HEADSTRESS$_{\text{Affix}}$ | ALIGN-L(Ictus, PrWd) | IDENT-OO(Ictus, acc.sg.) |
|---|---|---|---|---|---|---|
| a. | 9.5 | pádvatas | 4 | 4 | | 1.5 |
| b. | ☞ 5 | padvátas | 4 | | 1 | |
| c. | 7.5 | padvatás | | 4 | 1 × 2 | 1.5 |

In turn, the additional violations from IDENT-OO(Ictus, acc.sg.) help to exclude ictus mobility in forms with multiple underlying accents, but where neither of which is the head.

(48) gen.sg. *gómatas* 'having cows' ← /gáv-mant-ás/

| gáv-mant-ás (BASE: *gómantas*) | | | MAX-IO(Accent) | HEADSTRESS$_{\text{Affix}}$ | IDENT-OO(Ictus, acc.sg.) | ALIGN-L(Ictus, PrWd) |
|---|---|---|---|---|---|---|
| a. | ☞ 8 | gómatas | 4 | 4 | | |
| b. | 10.5 | gomátas | 4 × 2 | | 1.5 | 1 |
| c. | 13 | gomatás | 4 | 4 | 1.5 × 2 | 1 × 2 |

In cases where a derivational suffix introduces an accent, however, the accent of that suffix will surface as the ictus, even in cases of an accented root. Candidate a. here incurs violations of both MAX-IO(Accent) and HEADSTRESS$_{\text{Affix}}$, and hence is less harmonic.

(49) acc.sg. PPP *taṣṭám* 'fashioned' ← /tákṣ-tá-m/

| tákṣ-tá-m | | | MAX-IO(Accent) | HEADSTRESS$_{\text{Affix}}$ | IDENT-OO(Ictus, acc.sg.) | ALIGN-L(Ictus, PrWd) |
|---|---|---|---|---|---|---|
| a. | 8 | táṣṭam | 4 | 4 | | |
| b. | ☞ 5 | taṣṭám | 4 | | | 1 |

Although it is difficult to establish whether stems with at least two syllables in a monomorphemic stem and a consistent initial syllable ictus contain an inherent accent or have that initial syllable ictus to ALIGN-L(Ictus, PrWd) in the direct cases (nom.sg., acc.sg., etc.), and maintain the ictus in the same position by output-output faithfulness in the oblique cases

| pad-vant-ás (BASE: *padvántam*) | | MAX-IO | HEADSTRESS | IDENT-OO | ALIGN-L |
|---|---|---|---|---|---|
| a. | pádvatas | * | *! | | |
| b. ☞ | padvátas | * | | | * |
| c. | padvatás | | * | * | *!* |

(gen.sg., loc.sg., etc.), we should allow for the possibility. Let us assume for the moment that the neuter *s*-stem *manas-* 'mind' is treated holistically as a stem /manas-/. The nom./acc.sg. *mánas* then best satisfies ALIGN-L(Ictus, PrWd).[36]

(50) nom./acc.sg *mánas* 'mind' ← /manas-/

| manas- | | MAX-IO(Accent) | HEADSTRESS$_{\text{Affix}}$ | IDENT-OO(Ictus, acc.sg.) | ALIGN-L(Ictus, PrWd) |
|---|---|---|---|---|---|
| a. ☞ 0 | mánas | | | | |
| b. 1 | manás | | | | 1 |

In combination with an accented inflectional ending, such as gen.sg. /-ás/, the position of the ictus remains the same, because the violations of ALIGN-L(Ictus, PrWd) and IDENT-OO(Ictus, acc.sg.) are more severe than the violation of MAX-IO(Accent).

(51) gen.sg. *mánasas* 'mind' ← /manas-ás/

| manas-ás (BASE: *mánas*) | | MAX-IO(Accent) | HEADSTRESS$_{\text{Affix}}$ | IDENT-OO(Ictus, acc.sg.) | ALIGN-L(Ictus, PrWd) |
|---|---|---|---|---|---|
| a. ☞ 4 | mánasas | 4 | | | |
| b. 6.5 | manásas | 4 | | 1.5 | 1 |
| c. 5 | manasás | | | $1.5 \times 2$ | $1 \times 2$ |

Thus, besides accounting for evident exceptions to the Oxytone Rule as described in the preceding section, the present analysis of Sanskrit ictus assignment in terms of headedness has two decided benefits:

- There is no need to posit a lexical distinction between dominant and recessive morphemes – apparent dominance effects are reduced to a more general principle, headedness.
- There is no cyclic application of accenting and de-accenting processes that must apply. The correct results can be obtained from a single set of constraints that evaluate complete inputs. In short, the analysis is possible in a single stratum of a Harmonic Grammar, and achieves as much or more descriptive coverage than an analysis in Stratal Optimality Theory. Nevertheless, the model developed here should be extended to further data in order to systematically test its empirical validity.

[36] The morphologically parsed option that explains the form *mánas* would probably need to regard the suffix as pre-accenting (i.e., unaccentable) /-ˊas/, because neuter *s*-stems in Skt. almost invariably show the ictus on the initial (root) syllable (I know of one exception in the RV, $1 \times$ *tveṣás-* 'drive').

### 7.2.4 Greek HeadFaith

Within Vedic, another ultimate benefit of the use of HeadStress in the context of a harmonic grammar is that it may render the use of a HeadFaith constraint superfluous: apparent HeadFaith effects, such as persistent ictus on the derivational suffix *-tá-*, can fall out from the ganging of Max-IO(Accent) and HeadStress. For its part, however, it is not clear that Greek has any active HeadStress constraint. In brief, this is because Greek lacks any derivational suffixes that behave like that Ved. /-mant-/ or /-vant-/, i.e., unquestionably productive derivational suffixes that do not exhibit a consistent pattern of accentuation. Instead, morphologically complex forms in Greek can be essentially divided into three groups, from the accentual point of view:

1. a consistent lexical accent on the syllable corresponding to the suffix for all lexemes in the category., e.g., -τύ- [-tú-].

2. consistent recessive accentuation, e.g., for σι-/-τι [si-/ti-]stems.

3. some degree of variability across lexemes having the same derivational suffix, in which some have a lexical accent on the suffix, while others are recessive, e.g., nouns in -ρό-/-ρο- [-ró-]/[-ro-]: contrast ὄμβρος [ómbros] 'storm' with ξυρός [ksurós] 'razor'.

The third type has been extensively investigated by Probert 2006b, and I will discuss it in the following section. Types 1. and 2. were mentioned above under section 7.1.3, with examples at (24). The most efficient means of explaining the paradigmatically consistent accentuation of types 1. and 2. is through the use of a highly ranked/weighted HeadFaith constraint (see the definition at example (28) and compare the Japanese suffix /-(y)ó́/ at example (32)). Type 1 affixes will be accented morphemes, in which the high tone that they sponsor will define where the ictus must fall; type 2 affixes will be unaccentable morphemes, which appear to force the ictus outside of their domain.[37] Under the analysis proposed by Revithiadou (1999: 204–8; 222–4) for unaccentable derivational morphemes in Modern Greek and Russian, such morphemes sponsor a "floating" accent, which is required to be realized on the surface by HeadFaith, but the actual attachment of which to segments is determined by the default prosodic constraints of the language. The surface effect of an unaccentable morpheme is thus empirically indistinguishable from the deletion of all underlying accents.

As an example of the behavior of an unaccentable derivational morpheme in Ancient Greek, consider the form βασίλεια [basíleia] 'queen', derived from βασιλεύς [basileús] 'king'. The latter form evidently contains the accented derivational suffix /-eú-/, which consistently shows the high tone on the suffix to satisfy HeadFaith. The accent in the derivative, built with the suffix /-i̲a̲-/ (underlining here indicates unaccentability), exhibits recessive accentuation: nom.sg. βασίλεια [basíleia], (Ionic) gen.sg. βασιλείης [basileíɛ:s].[38] The UR of [basíleia]

[37] In a language with highly ranked HeadFaith, but low-ranked HeadStress, there could be unaccented derivational morphemes that give the appearence of simply accepting the position of the ictus found in their bases. Ancient Greek does not appear to have possessed any such derivational morphemes, or, if it did, the effects of the Law of Limitation may render them indistinct from unaccentable morphemes.

[38] Attic gen.sg. βασιλείας [basileía:s].

thus has two underlying accents, one linked to the morpheme /-éu-/, the other a floating accent sponsored by /-ia-/. Because /-eú-/ is not the head, the faithful realization of its accent does not satisfy HeadFaith (cf. candidate b. in the tableau below); since the floating accent sponsored by /-ia-/ may be linked anywhere and satisfy HeadFaith, the realization of the ictus in a position other than the position desired by Recessive Accent is gratuitous (cf. candidates c. and d. below). Hence, candidate a. below, which violates only Max-IO(Accent), by failing to realize the accent of /-eú-/, is optimal.

(52)

| basil-eú-ia | | Law of Limitation | HeadFaith | Max-IO(Accent) | Recessive Accent |
|---|---|---|---|---|---|
| a. ☞ | basíleia | | | * | |
| b. | basiléia | | *! | | * |
| c. | basiléia | | | * | *! |
| d. | basileiá | | | * | *! |
| e. | básileia | *! | | | |

Accentuation in morphologically complex stems of Greek can thus result straightforwardly from highly ranked HeadFaith. One overall practical wrinkle, however, is that while morphological categories of type 1. must clearly have a lexical accent, categories belonging to type 2. are inherently ambiguous: all lexemes belonging to the category etymologically might have morphologically simplex stems, and therefore have recessive accentuation because no affixal head attempts to impose its accent. This possibility is particularly likely in the case of forms derived etymologically with primary derivational suffixes;[39] such forms are built largely to verbal roots, and often lack a base which might have been accented differently.

Up to this point, I have advanced the argument that the accentual behaviors of Greek and Sanskrit in morphologically complex words can result transparently by reference to morphological heads.[40] A theoretical benefit this model confers is to exclude the need for any

[39] Cf. 4.1 above on the distinction between primary and secondary derivational suffixes.

[40] I have not touched whatsoever on the accentuation of compounds in either language. In short, I believe that the accentuation of compounds in both languages is subject to a head-based analysis as well, though the relevant condition for headedness appears not to be semantic headedness (as in Thompson Salish or Japanese compounds) but syntactic headedness. In the earliest Vedic, compounds that reflect a head–complement relationship are subject to strict HeadStress effects: the larger class of determinative compounds (traditionally called *tatpuruṣa*), including probably all with a verbal second member (type *havir-ád-* 'eating the oblation'), as well as verbal governing compounds, assign the ictus to the head; compounds that reflect a head–specifier relationship (essentially determinative compounds traditionally called *karmadhāraya* compounds), or which are exocentric (and thus headless, i.e., *bahuvrīhi* compounds), simply select the leftmost underlying accent.

Complicating this distribution is the fact that, as Kiparsky (2010: 175–6) has pointed out, derivational suffixes may attach outside of two compounded elements, and as a head, induce a dominance effect. For example, a form such as *mahādhaná-* (10× RV) 'great stakes', traditionally described a *karmadhāraya*, has the structure [ [ [ mahā́ ] [ dhan ] ] á ], in which /-á-/ takes scope over the compounded elements, and thus takes the ictus

derivational layers in the phonological architecture. However, classical Attic Greek presents a number of forms, especially among so-called "contract verbs", that regularly exhibit an accentuation that would imply the assignment of an accent prior to the contraction of underlying vowels. For instance, the verb φιλέω [pʰiléɔː] 'love' shows systematic contraction of the stem-final vowel with a vowel-initial inflectional ending; e.g., 3.pl.pres. behaves as if /pʰile-oːsi/ → /(pʰilé)-(*oː)si/ (footing and accent assignment) → [pʰilôːsi] (vowel contraction and output). If the recessive accent were assigned on an underlyingly contracted form /pʰiloːsi/, the result would be [X]pʰí(*loː)si]. Noyer (1997) and Kiparsky (2003) adduce evidence of this sort to argue for the need for some derivational layer in Attic Greek phonology; such analyses would effectively recapitulate the diachrony of vowel contraction in the syncrhonic phonology. The issue is thoroughly surveyed by Probert (2010), who considers further evidence in which operation with underlying uncontracted vowels produces incorrect results, and who ultimately cautiously endorses a non-derivational model with greater burden on morphology and the lexicon to account for the historical effects of contraction. A similar survey of the issues involved would take us too far afield from the essential generalization concerning accentuation, and in particular the relation to derivational morphology with which I am concerned here, and I therefore refer the reader to Probert's discussion.

## 7.3 Probert on Frequency Effects and Lexical Accentuation

With functioning predictive models of accentuation in morphologically complex words in Greek and Vedic in hand, we can now turn to the question that opened this chapter: does any demonstrable relation hold between the productivity of a morphological process and the accentuation of its instantiating members? First, I should clarify why one might expect for productivity to play some role in the prosody of languages with lexical accent systems. *Ex hypothesi*, morphemes are supposed to be equipped with prosodic information, but for that prosodic information to enter the derivation and be assessed by the phonology, the presence of the morpheme must be recognized by the speaker. If a lexeme is produced or processed holistically (or at least, without complete morphemic decomposition), the prosody of that lexeme cannot be attributed to the property of some morpheme, although that lexeme might itself have a lexical entry with the relevant prosody stored. The problem is essentially the

by HeadStress. Hence, it is crucial first to identify and separate these synthetic compounds (which also exist in Greek) from true compounds of the structure only [ [ X ] [ Y ] ]. The traditional description of accentuation in karmadhārayas, and apparently productive accentuation the right edge of the stem, may follow from a generalization based on accentuation in synthetic karmadhārayas. It may be most appropriate to say that the determination of compound accentuation shifts from a syntactic determination to a semantic determination; thus, all determinative (tatpuruṣa and karmadhāraya) compounds being endocentric, the semantic heads tend to attract the ictus.

Similarly, in Greek, compounds that reflect a head–complement relationship exhibit effects of HeadFaith, whereas compounds that reflect a head–specifier relationship have recessive accentuation. However, because the prosodic grammar relies on HeadFaith, not HeadStress, compounds that satisfy the head-complement relation, but whose second members are either unaccented or themselves headed by unaccentable morphemes, will give the appearence of recessive accentuation as well. Whether the suggestions here concerning Vedic and Greek compound accentuation are entirely correct remains to be thoroughly investigated.

same as the problems posed for the analyst at 7.1.6 above: does surface [ūtáye] dat.sg. 'aid' ($100\times$ RV) derive directly from a complete lexical entry /ūtáye/, or from a parsed /ū-ti-é/, with ictus assigned to the head /-ti-/ to satisfy HEADSTRESS in Vedic?[41] The very absence of morphological structure, as potentially diagnosable by measures of productivity, in turn, has the potential to account for both changes in prosodic behavior of morphological classes, and seeming irregularities in the prosody of individual lexemes. Revithiadou (1999: 223, fn. 31) notes: "It is well-known that often loss of morphological boundaries causes a chain of changes which can have an effect on the prosodic structure of the word as well."

By far the most thorough attempt to establish some relation between prosody and the morphological structure of individual lexemes of which I am aware, concerning any older Indo-European language, is Probert (2006b)'s survey of accentuation in Greek nominal and adjectival categories built with thematic derivational suffixes (e.g., -μο- [-mo-], -νο- [-no-], -ρο- [-ro-], -το- [-to-], -λο- [-lo-]). This study wisely concentrates on categories and subcategories that meet two conditions: first, all of these suffixes sometimes show the high tone on the suffix, thereby indicating that they are to be analyzed as accented suffixes; second, the members of the categories are divided in accentuation (e.g., among $46$ types using the suffix -ρο- [-ro-] to build nouns, $15$ show the high tone on the suffix -ρό- [-ró-], while the other $31$ show recessive accentuation). Furthermore, because Greek has the ranking HEADFAITH $\gg$ FAITH,[42] the recessive accentuation of a word such as βόθρος [bót$^{\text{h}}$ros] 'hole, trench' cannot be attributed to a lexical accent on the root; i.e., a UR /bót$^{\text{h}}$-ró-s/ would be predicted to select $^{\text{X}}$[bot$^{\text{h}}$rós] as the output, not actual [bót$^{\text{h}}$ros]. The existence of βόθρος [bót$^{\text{h}}$ros] alongside forms with non-recessive accentuation such as ξυρός [ksurós] 'razor', built with ostensibly the same derivational suffix, requires a general account that can apply to specific lexemes.

The essence of Probert's account is that forms such as βόθρος [bót$^{\text{h}}$ros] are "demorphologized", and subject to the phonologically default accentuation of the language in question (for Greek, the recessive accent). Probert (2006b: 291) summarizes:

> When a word has undergone 'demorphologization' its accentuation can no longer be determined by the presence of an inherently accented suffix as the suffix is no longer treated synchronically as present. The word may retain its non-recessive accentuation but the necessary accentual property now becomes a characteristic of the whole—synchronically unanalysed—stem. On the other hand, the word may lose its inherent accent altogether, in which case a recessive accent will be assigned by default.

[41] The derivational suffix /-ti-/ must be unaccented in Vedic, as demonstrated by derivatives in /-mant-/ from nouns built with /-ti-/: the RV has *puṣṭimánt-* 'rich in food' (cf. *puṣṭí-* 'growth; food'), *śruṣṭimánt-* 'having willingness' (*śruṣṭí-* 'willingness'), *ɻṣṭimánt-* 'having a spear' (cf. *ɻṣṭí-* 'spear'), *bhɻṣṭimánt-* 'having a point' (cf. *bɻṣṭí-* 'point, corner'), *vɻṣṭimánt-* 'containing rain' (cf. *vɻṣṭí-* 'rain'), *svastimánt-* 'having well-being' (cf. *svastí-* 'well-being'). Both my analysis at 7.2.3 and Kiparsky (2010)'s analysis would require that the suffix /-ti/ not possess any accent to account for those *mant*-stems, which in turn entails that the base nouns to which those *mant*-stems are derived in Vedic must show the ictus on the suffix [-tí-] due to HEADSTRESS or the Oxytone Rule.

[42] This ranking is evident from forms such as gen.pl. ἀστερίσκων [asterískɔ:n] ← /aster-ísk-ô:n/, in which the accent of the head wins over the accent of the inflectional suffix.

Probert thus considers that a "demorphologized" lexeme has two potential prosodic fates: store a lexical accent that would render a surface form equivalent to an older morphologically parsed version, or treat the stem as unaccented altogether, and produce a form with the phonological default.

In the course of the 2006 book, Probert principally looks to token frequency of individual Greek lexemes as an explanatory factor. 7.3.1 briefly reviews some of Probert's results in that work; I find that, although most of Probert's interpretations of the facts are reasonable from the psycholinguistic point of view, absolute token frequency alone is not a completely adequate predictor of accentuation within those categories. I then move to my own examination of some Greek and Vedic nominal categories at 7.4.

### 7.3.1 Token Frequency

Probert (2006b: 292) states an overall conclusion for the role of token frequency[43] as a determinant of accentuation in Greek: very high token frequency and very low token frequency forms are "most likely" to exhibit a lexical accent, whereas lexemes of moderate frequency are more likely to show the default phonological accent of Ancient Greek. Granting, for the moment, the accuracy of the observation that lexical and default accents correlate with different frequency ranges, the classification of lexemes into LEXICAL ACCENT and DEFAULT ACCENT bins takes on a U-shaped (or parabolic) distribution: very high and very low frequency fall into one bin, ranges in between in the other bin.[44]

The psycholinguistic literature on token frequency effects, surveyed at 3.2.1 above, generally supports Probert's observations and interpretation. Precisely because high token frequency forms are subject to faster lexical retrieval (presumably because they are produced and processed as full forms), they may preserve and maintain a lexical accent, as part of a stem or whole word, which conflicts with the language's default phonological accent. Conversely, the lowest token frequency lexemes are those most likely to be accessed via a morphologically parsed route, and thus if they contain an affix that has a lexical accent, they are likely to allow the lexical accent of the affix to emerge. For lexemes of middling token frequency, the phonologically preferred means of prosodic prominence is presumed to emerge because those lexemes are too frequent to undergo parsed access with regularity, but not frequent enough to be stored as a stem or full form with a lexical accent.

It is somewhat surprising to claim that, on the one hand, a form such as βόθρος [bót$^{h}$ros] 'hole, trench' ($31\times$ in Probert's corpus of Greek) should be regularly accessed from memory as a full stem /bot$^{h}$ro-/, and for that reason subject to "demorphologization", yet not have been stored in memory with the (presumed) lexical accent of the suffix /-ró-/, which, at the

[43] Token frequency in Probert 2006b always means the absolute token frequency of a given lemma, obtained from a Greek corpus of $\sim$ 3400000 words, drawn from authors from Homer into the 1st c. BCE. One general worry about Probert's approach is that it may collapse data from several different layers of the language.

[44] If the distribution were indeed neatly parabolic, then one should find that a regression model with a quadratic term (in this case, token frequency and squared token frequency) should work well to predict the data, and the regression terms based on token frequency should be significant predictors. We will see that, for Probert's data on Greek [ro-]stem nouns, this is indeed the case.

birth of the lexeme, should have led to the generation of a surface form with suffixal accent, *[bot$^{h}$ró-].[45] Let us assume that, at some point, indeed a stem /bot$^{h}$ró-/ existed in the lexicons of Greek speakers; apart from the position of the high tone, /bot$^{h}$ró-/ would not appear to violate any active phonological constraints of Greek. At the level of the individual lexeme, "regularization" of the accent might occur when the USE-LISTED constraint (cf. Zuraw 2000: 49–53) indexed to that particular lexeme is weaker than the general phonological constraints governing word-level prosody.[46] Clearly, USELISTED$_{[\text{bot}^{h}\text{ró-}]}$ would outrank, say, *[b], which might cause /bot$^{h}$ro-/ to be output as $^{X}$[pot$^{h}$ró-]. If, however, RECESSIVE ACCENT were to outrank USELISTED$_{[\text{bot}^{h}\text{ró-}]}$, the output would be the attested [bót$^{h}$ro-]. In short, because the language has active constraints that attempt to enforce a given word-level prosody, substantial evidence would be required on the part of the learner to accept a stem in the lexicon with some "irregular" prosodic feature.

While Probert's theory of "demorphologization" thus appears psycholinguistically sensible, and aspects of its predictions can be operationalized in terms of USE-LISTED constraints, an outstanding question is whether Probert's observed tendencies stand up to closer statistical scrutiny. As a test case, I consider the data on Greek nouns derived with the suffix -ρο- [-ro-] (Probert 2006b: Ch. 6, esp. 170–1). Probert lists $46$ distinct lexemes (types) collected from lexicographical resources, and token frequencies derived from a corpus of $\sim 3400000$ tokens. $14$ of the types did not occur in Probert's corpus, and Probert regards those items with token frequency $0$ as the lowest token frequency items.[47] For lexemes with a token frequency above $0$, Probert sorts the items into ranges of $100$ tokens (i.e., words with $1$–$100$ tokens in the corpus, etc.).[48] Probert draws the conclusion of a U-shaped relation between frequency and accentuation in [ro-]stems on the basis of the fact that more lexemes with recessive accentuation fall into the $1$–$100$ range than occur $0\times$ ($16$ vs. $9$), and among the four items occurring more than $500$ times, three are accented on the derivational suffix (ἱερόν [hierón] 'offering; temple', ἐχθρός [ek$^{h}$t$^{h}$rós] 'enemy', νεκρός [nekrós] 'corpse') and one is recessively accented (δῶρον [dɔ̂:ron] 'gift'). If one takes seriously Probert's division between lexemes unattested in her corpus, and those attested $1$–$100$ times, a $\chi^2$-test suggests a non-significant difference between the number of lexically accented and the number of reces-

[45] The possible role of substantivization or concretizations in such seeming accent retractions is discussed below. βόθρος in particular is also treated in Peters 1999. Neri (2011: 230–3) also treats the accentuation of this form as due to substantivization, though he assumes the presence of an original suffix /-d$^{h}$ró-/.

[46] Similarly, we can think of the inherent accents pertaining to roots and affixes as projecting USELISTED constraints in proportion to the number of parsed tokens of that morpheme. In case the USELISTED constraint indexed to a particular morpheme falls below the weight of general phonological constraints governing word-level prosody, we should expect to find the possibility of prosodic regularization for all members of that morphological category. The lexical accents of roots and affixes are then likely bound up with the productivity of an affix and the family size of a root (see 3.2.5 above on family size effects).

[47] However, it is clear that some token frequencies could have been radically different, had Probert made slightly different choices of corpus. For instance, κύλινδρος 'cylinder' does not occur at all in Probert's corpus, but had her corpus included some texts of the Pythagorean corpus, it would have occurred with fairly high frequency.

[48] The $1$–$100$ range really ought to be broken down more finely, since a substantial psychological difference is to be expected for a word found only once (e.g., ἵππερος [híp:eros] 'horse-fever') and one found eighty times (e.g., λάφῡρα [lap$^{h}$u:ra] 'spoils of war').

sively accented lexemes.[49] The barplot in Figure 7.1 shows the distribution of recessive and lexical accent among [ro-]stems in Probert's corpus (essentially reproducing the figure on Probert 2006b: 172).

Treating token frequency as a predictor of accentuation (i.e., binary classification as either lexically accented or recessively accented) in a regression model, on the other hand, suggests that token frequency can be a useful predictor.[50] Given the possible U-shaped distribution of the outcomes, linear regression with the token data will not be an effective predictor.[51] Adding a quadratic term (i.e., `token frequency`$^2$), however, should be much more effective if the underlying distribution is indeed parabolic in form. Indeed, both `token frequency` and `token frequency`$^2$ then appear to be significant predictors.[52] This quadratic regression model correctly predicts the accentuation for $\sim$ 72% of forms $(33/46)$.[53] Excluding an intercept term does not improve the model's overall accuracy, but it does make the probabilities assigned in one direction or another stronger, and thus the correct items are more robustly correct without an intercept term. The greatest source of inaccurate predictions here are the lower-frequency lexemes with lexical accent; in the main, the model sees higher token frequency as a better indicator of lexical accentuation. For frequency above $\sim$ 250, the model predicts accent on the suffix -ρό- [-ró-]. We can confirm that a quadratic model of token frequency effects is on the right track, because it performs substantially better than a linear model (which has an accuracy of only 65%), and a model with a cubic term (i.e., `token frequency`$^3$) performs only marginally better (accuracy of 74%), but at the cost of rendering all of the model's predictor terms non-significant.

Based on these results, we have reason to believe that token frequency may indeed substantially affect a lexeme's prosody in Greek. However, given the substantial residue of recessively accented forms with low token frequency that cannot be accurately predicted on the basis of token frequency alone, a better model would surely include more factors. Probert does, at one point, give an indication that productivity more generally could play a role in the accentual behavior of a class. Of nouns derived with the suffix -λο- [-lo-] (excluding -ιλο- and -υλο-), Probert (2006b: 236) states: "The massive incidence of recessive accentuation among nouns with -λο-... is, I think, a result of the general lack of productivity and synchronic viability of the suffix -λο-... The suffix -λο- was not productive in Greek and was not readily recognized as a morphological element in a word, especially if that word was a noun." Indeed, nouns derived with [-ro-] would appear to have no or marginal productivity as well, only

[49] $\chi^2 = 0.1458$, $p = 0.7025$, but because the sample size is so small (a total of 14 words attested $0\times$ and 21 attested $1$–$100\times$), the test may be inaccurate.

[50] For the four lexemes attested more than $500\times$ in her corpus, Probert does not give complete token counts. For reasons described in Probert 2006b: 168, fn. 20, precisely reproducing Probert's corpus is not possible. In order to obtain a more exact count of those four lexemes, I used the Perseus Project (`www.perseus.tufts.edu`) to build a corpus of approximately the same size, drawing on the same authors as did Probert, and collected the token frequencies from that corpus.

[51] Indeed, such a linear model finds token frequency to be a non-significant predictor ($Pr(> |t|) = 0.530238$) of accentuation.

[52] $Pr(> |t|) = 0.03901$ and $0.04881$, respectively.

[53] The model assigns a probability of having the ictus on the derivational suffix ,thus [-ró-]. I counted a probability greater than .5 as predicting that lexical accent, and a probability less than .5 as predicting recessive accentuation.

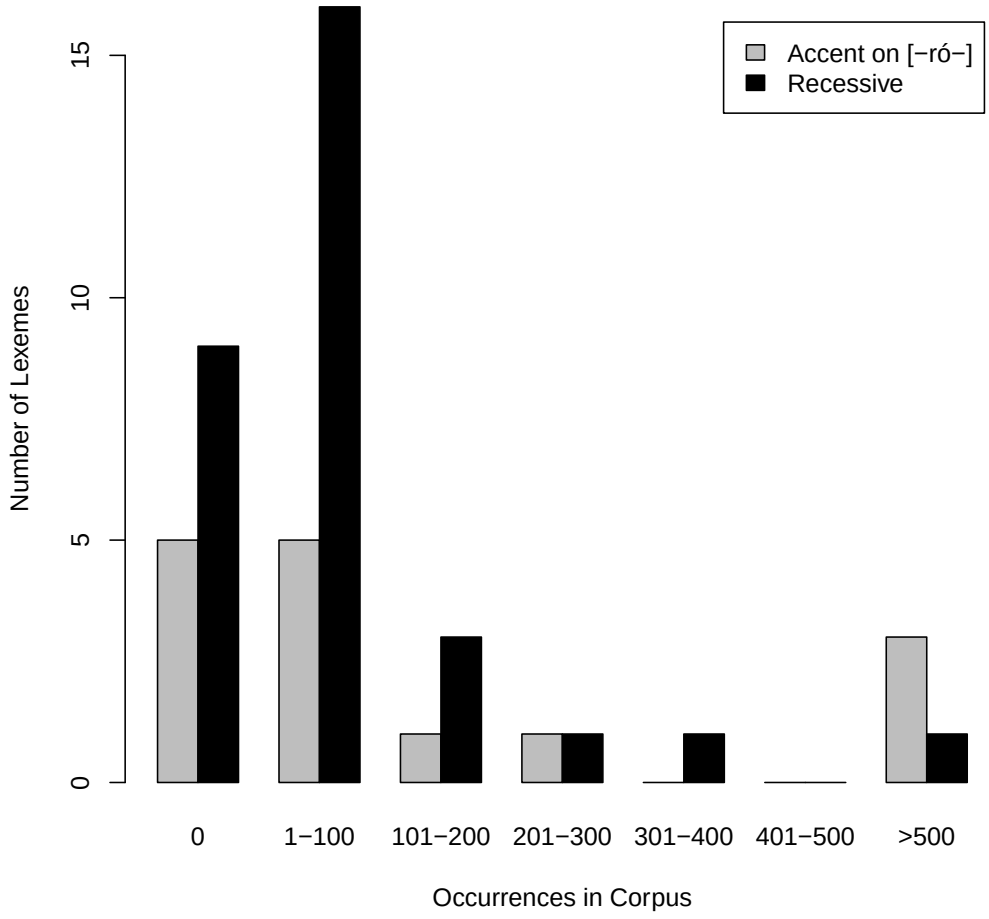



Figure 7.1: Frequency Distribution of Recessive and Lexically Accented [ro-]stems in Greek (after Probert 2006b: 170–2)

4 hapax legomena occurring among an estimated 6320 tokens of the category in Probert's corpus. Perhaps low productivity is a necessary condition for "demorphologization" in the first place.

### 7.3.2 Substantivization

Also worth noting is Probert (2006a), where the author extends a similar treatment to feminine nouns in -η [-ɛː], but with greater attention to semantics as a determinant of accentuation. In brief, the development of more "concrete" semantics among that class of nouns appears to correlate with recessive accentuation. For example, the root of Gk. σκάφω [skáp$^{h}$ɔː] 'dig' attests two feminine *ā*-stems, each with different accentuation: we find σκαφή [skap$^{h}$ɛ́ː] 'digging' and σκάφη [skáp$^{h}$ɛː] 'trough; light boat' – the former refers to the act of digging in an abstract sense, whereas the latter identifies a concrete object in the world. Forms with abstract semantics that derive directly from a base verb may be transparent, productive derivatives; conversely, in cases of semantic drift away from an etymologically original base, parsed access may become impossible, if the form does not seem to contain its base. Probert's observations here on Greek forms appear to be entirely in accord with the results on polysemy obtained by Hay (2003: Ch. 5) for English.

Overall, I believe that the theory of "demorphologization" has many theoretical merits in itself, and parts of the specific form in which Probert advances it accord nicely with the basic and often-replicated psycholinguistic results concerning whole-word/lemma token frequency. Furthermore, more rigorous analysis on a subset of Probert's data appears to partially bear out her conclusions. Granting that "demorphologization" can indeed account for variable ictus assignment among words of the same morphological category, the problem is then simply to construct more powerful models based on further kinds of data. The following section, which exmines several nominal categories in Homer and the *Ṛgveda*, aims to act as a step in that direction.

## 7.4 Productivity, Parsability, and Accentuation

The principal objective of this section is to present statistics on the quantitatively assessed productivity of several nominal categories in Greek and Vedic, based on their distributions in Homer and the RV. Productivity in itself ought to have direct consequences for the corresponding impact of token frequency upon "demorphologization". Since a derivational morpheme is assumed to be more parsable and present among categories with high $\mathscr{P}$, lexemes that might fall into the dangerous middling token frequency range described above may remain sufficiently parsed for the accentuation of the derivational morpheme to be decisive; for unproductive categories, full-form or lemma storage and retrieval will be the rule for most forms, and most of the category's members might be subject to "demorphologization".[54] Af-

[54]Note also that base-derivative relative token frequency is probably a better predictor than absolute token frequency (Hay 2001; cf. 3.2.3 above), but deciding what the base of a primary derivative in Greek or Sanskrit might be is not a straightforward matter. Perhaps the principal verbal stem built to a given root for nouns with

ter presenting summary statistics on each category, I will comment on how the accentuation of the category as a whole and some specific forms ought to be interpreted in light of the general constraints on ictus assignment discussed in the preceding sections.

Let us then formulate some explicit hypotheses about the accentual behaviors to be encountered depending upon parsability as measured by $\mathscr{P}$. For a highly productive morphological category, we should expect to find absolutely consistent accentuation across all members of the category: the prosodic property of the derivational morpheme will always be decisive. In case of a wholly unproductive morphological category, the etymological derivational morpheme may not even have an independent lexical entry, with prosodic information particular to that morpheme; in such a case, the etymological members of the category may have either a lexical accent particular to the entire stem, or a show an accent determined by the phonology. Thus, unproductive categories would be predicted to show either: a) variable accent patterns in their members or b) perhaps all members could exhibit the phonologically determined accentuation.

Lists of the types belonging to each category treated here come from Risch 1974 and the index of Grassmann 1872 [1976]; token frequencies were collected from Lubotsky 1998 and WordHoard 2004–2011. In all cases, the frequency of occurrence of the simplex and its frequency in any compounds have been summed to produce the total token frequency for a given lemma. Some Homeric compounds not listed with their simplexes in Risch 1974 were identified by reference to van Strien-Gerritsen 1973.

### 7.4.1 Greek [si-]/[ti-] and [tú-]Stems

The two nominal categories represented by members such as βρῶσις [brɔ̂ːsis] 'food' or μάντις [mántis] 'prophet' on the one hand, and βρωτύς [brɔːtús] 'food' or κιθαριστύς [kit$^{h}$aristús] 'art of playing the *kithara*' on the other, I think merit a side-by-side comparison.[55] This is because the two categories share many functional and semantic properties, largely deriving feminine abstract nouns either from verbal roots or nouns. In Homer, however, their choice of bases differs somewhat: for [si-]/[ti-]stems, only derivatives directly to a verbal root (often in its zero-grade allomorph) are attested, whereas the [tú-]stems attest both derivatives from verbal roots (e.g., γραπτύς [graptús] 'scratching', from γράφω [grap$^{h}$ɔː] 'scratch; write') and nouns (e.g., κιθαριστύς [kit$^{h}$aristús] to κίθαρις [kít$^{h}$aris] 'lyre; a plucked string instrument' ). Most importantly, for the problem at hand, they differ in accentuation: all forms derived with [-tú-] (except for the neuter ἄστυ [ástu] 'town', which may be etymologically distinct) in Homer show the high tone on the derivational suffix, whereas all forms derived with [-si-]/[-ti-] exhibit recessive accentuation, and so, in practice, have the high tone on the root syllable.

The raw frequency and derived productivity statistics for these two categories are given

deverbative semantics is a sensible base, but how clear that base-derivative relationship might be, formally speaking, is hard to say.

[55] Though I transcribe the suffix as [-tú-] throughout here, as it was etymologically, the vowel was in fact long, at least in the nom.sg., as can be discerned metrically in Homer, thus [-túːs].

in Table 7.2.[56]

| Category | $V$ | $N$ | $n_1$ | $\mathscr{P}$ | $S$ | $\mathscr{I}$ |
|---|---|---|---|---|---|---|
| [-si]/[-ti-] | 46 | 422 | 9 | .02132701 | 50.48234 | 1.097442 |
| [-tú-] | 19 | 197 | 10 | .05076142 | 43.99002 | 2.315264 |

Table 7.2: Productivity Statistics for Homeric [si-]/[ti-] and [tú-]Stems

Taking the $\mathscr{P}$ of root and sigmatic aorists calculated in Chapter 5 as a baseline for (non-)productivity and parsability ($.0003656$ and $.0234108$), both categories here would appear to be productive, though the [tú-]stems clearly much more so than the [si-]/[ti-]stems. Given the similar number of hapax legomena in each category, the two appear to be of like degrees of usefulness, though the frequency distribution of [tú-]stems translates into a much higher degree of pragmatic potentiality ($\mathscr{I}$; cf. 2.2.2 and 2.2.3.1). In essence it seems that, although both categories could be labeled as 'productive', broadly speaking, [tú-]stems are indubitably parsable. Indeed, most of the tokens for the [tú-]stem category are contributed by ἄστυ [ástu] 'town' ($139\times$), which, on account of its different gender may not properly belong with the other [tú-]stems, in which case the $\mathscr{P}$ for the category would measure as an overwhelming $0.1724138$.[57]

Furthermore, while the [tú-] stems are almost always completely phonologically transparent with respect to their bases, [ti-]stems are not. For instance, ῥῆσις [rêːsis] 'speech' is explicable only as continuing an IE *[u̯r̥h₁tí-]; the verbal base, εἴρω [eirɔː] 'speak' $<$ *[u̯érh₁-i̯e/o-], is hardly synchronically transparent.[58] The very fact that Homeric [si-]/[ti-]stems are only found built directly to verbal roots, and largely do not attest a clear root formation, speaks further to the lower parsability of the [ti-]/[si-]stems generally.[59]

This distinction in productivity and parsability between the two categories under consideration here has the potential to account for the universal recessive accentuation of Greek [si-]/[ti-]stems. Historically, it is likely that most, if not all [si-]/[ti-]stems showed a surface ictus on the derivational suffix: most of the cognate category in the RV (see further below) indeed shows the ictus in that position, the usual zero grade of the root in both Greek and Vedic is suggestive of an ictus in that position, and cognate forms in Germanic (e.g., Old English

[56]For the sake of simplicity, I have directly followed Risch's classification of forms for these categories. Thus, although ἄστυ [ástu] is a neuter, I include it with the [tú-]stems, and although πόσις [pósis] 'lord' in origin is an *i*-stem to a root ending in [-t] (IE *[póti-]), it is included among the [si-]/[ti-]stems. The inclusion of compounds also has a substantial effect, because μῆτις [mêːtis] 'wisdom' occurs in the very common epithet of Odysseus, πολύμητις [polúmeːtis] 'of many wiles' ($89\times$). If all of the tokens of that epithet were excluded, the [si-]/[ti-]stems would have an accordingly higher $\mathscr{P}$ ($.02702703$), but which would remain notably less than $\mathscr{P}$ for the [tu-]stems.

[57]Without including the tokens of πόσις [pósis] 'lord', the $\mathscr{P}$ of the [si-]/[ti-]stems likewise appears higher, but still substantially less than for [tu-]stems, measuring as $.02601156$.

[58]Note, though that some verbal nouns, such as the passive aorist infinitive ῥηθῆναι [rɛːtʰêːnai], with the same root allomorph [rɛː-] as found in the [si-]stem, persists in use into the later history of Greek.

[59]Some items, such as βάσις [básis] 'stepping, going' or δόσις [dósis] 'giving', do stand alongside well-attested root aorists.

*mæd* < *[meh$_1$tí-] = Gk. μῆτις [mêːtis]) sometimes show the effects of Verner's law, likewise indicative of an ictus on the derivational suffix. What the comparatively lower $\mathscr{P}$ of Homeric [si-]/[ti-]stems might suggest is that they were perhaps insufficiently parsable, and have been systematically "demorphologized" as a class. With the exception of μῆτις [mêːtis] 'wisdom' (which is probably overrepresented in this corpus) and πόσις [pósis] 'lord', no Homeric [si-]/[ti-]stems are of especially high token frequency, and hence the absence of any lexemes that preserve an accent on the derivational suffix is unsurprising.[60] Arguing against this interpretation, however, is the simple fact that with respect to the baseline measurement of productivity offered by the sigmatic aorists, we cannot quantitatively define [si-]/[ti-]stems as unproductive.

The different histories of [-tú-] and [-si-]/[-ti-] in the history of Greek is worth comment as well. To judge from the number of types attested in each category, [-si-]/[-ti-] was enormously useful, attesting well over 5000 distinct types (Buck and Petersen 1949: 574); the fate of the [tú-]stems, was the converse, attesting only about 115 types, mostly in works before the end of the Classical period, and otherwise in lexicographical works (e.g., Hesychius). This history is hardly to be predicted on the Homeric attestation of these two categories. The fact that the numerous hapax legomena [si-]/[ti-]stems are absolutely consistent in showing recessive accentuation, indicates that the suffix was acquired, reasonably, as unaccentable /-si̲/ (cf. the discussion at 7.2.4 above). An account of why [-tú-] and [-si-]/[-ti-] go on to such radically different futures is a question deserving of an investigation in itself.

### 7.4.2 Greek [ma-] and [mon-]Stems

Let us now look to two other noun-forming suffixes that, like the [si-]/[ti-]stems, usually exhibit recessive accentuation. Among abstract neuter nouns formed with the suffix -μα- [-ma-][61] (e.g., νόημα [nóɛːma] 'perception, thought') in Homer, recessive accentuation appears to be exceptionless; among masculines and adjectives built with the suffix -μον- [-mon-] (nom.sg. -μων [-mɔːn]), recessive accentuation is typical, but Homer knows some eight nouns that instead have the high tone on the derivational suffix. All of the forms with high tone on the suffix are nouns, and most of those with recessive accent are adjectives.

From the historical point of view, the recessive accent of the neuters accords with the initial-syllable ictus typical of Vedic neuter *man*-stems, but instances of recessively accented [mon-]stems disagree with the usual ictus on the derivational suffix of Vedic masculine *mán*-stems (see further below).[62] The [mon-]stems are particularly common as second members

[60] Complicating this interpretation, however, is the fact that Vedic *ti*-stems appear not to have an underlying accent; cf. fn. 36 above. If PIE possessed highly ranked HeadStress like Vedic, then the recessive accentuation of Greek [si-]/[ti-]stems might simply be attributed to a change from a HeadStress to a HeadFaith system. However, if Proto-Greek (?) learners were exposed to numerous forms such as *[dotí-] or *[g$^{\text{w}}$m̥tí-], and learned high-ranking HeadFaith, the reasonable analysis would be to attribute the ictus to an accented morpheme /-tí/, which then makes a "demorphologization" account necessary.

[61] [-ma-] here historically continues *[-mn̥].

[62] Much like the Vedic *ti*-stems, however, the evidence from derived adjectives in /-vant-/ suggests that the masculine-forming suffix /-man-/ is not accented in Vedic, but gets its ictus on account of HeadStress: compare *ātmán-* 'body, self' alongside *ātmanvánt-*, and *dhvasmán-* 'smoke cloud' alongside *dhvasmanvánt-*.

of compounds (indeed, many of those in Homer appear solely in compounds), and Risch (1974: 52–3) in fact supports the view that the Greek adjectival [mon]-stems do not derive directly from verbal roots, but are abstracted from adjectival possessive compounds made from original neuters in [-ma-]. Thus, νόημων [nóɛːmɔːn] 'thoughtful' would be abstracted from the compound ἄνοημων [ánoɛːmɔːn] 'without thought, thoughtless' ← neut. νόημα [nóɛːma] 'perception, thought'.

Table 7.3 gives the productivity statistics pertaining to these categories. The tokens of [mon-]stems, may, however, be seriously overrepresented on account of the very frequent personal name Ἀγαμέμνων [agamémnɔːn] $(184\times)$.[63] If these tokens are excluded, $\mathscr{P}$ for the [mon-]stems instead measures as $.02465753$.[64] At the same time, the personal names Πάμμων [pámːɔːn] and Ἐχέμμων [ekʰémːɔːn] contribute two hapax legomena to the category.

| Category | $V$ | $N$ | $n_1$ | $\mathscr{P}$ | $S$ | $\mathscr{I}$ |
|---|---|---|---|---|---|---|
| [-ma-] | 66 | 1108 | 20 | .01805054 | 95.4988 | 1.446952 |
| [-mon-] | 35 | 549 | 9 | .016393442 | 43.77885 | 1.250824 |
| [-mon-] adj. | 19 | 377 | 5 | .0132626 | 23.66278 | 1.245409 |
| [-mon-] noun | 16 | 172 | 4 | .02325581 | 20.45442 | 1.278401 |
| [-mon-] noun w/o PN | 14 | 170 | 2 | .01176471 | – | – |

Table 7.3: Productivity Statistics for Homeric [ma-] and [mon-]Stems

The parsability of both the neuters in [-ma-] and masculines/adjectives in [-mon-] appears to be lower than in the case of the [si-]/[ti-]stems treated above. If it was correct to say that [si-]/[ti-]stems universally exhibit recessive accentuation because the parsability of the category at the Homeric period was too low, then we should expect recessive accentuation for all [ma-] and [mon-]stems as well.[65] This explanation suits the [ma-]stems and [mon-]stem adjectives perfectly well, but not the [mon-]stem nouns, where $8/16$ types show an accented derivational suffix. The higher $\mathscr{I}$ of the [mon-]stems with respect to the [si-]/[ti-]stems ($1.250824$ vs. $1.097442$) suggests the presence of a larger proportion of low frequency types in the population of [mon-]stems, which could point to more regular frequent parsing, and so admit of some lexically accented forms.

Since the [mon-]stem nouns in fact show variability between recessive accentuation (e.g., δαίμων [daímɔːn] 'spirit') and high tone on the suffix (e.g., λειμών [leimɔ́ːn] 'meadow'), a look into the relation between lemma token frequency and accentuation in this subtype,

[63] In the subdivision of [mon-]stems into adjectives and nouns, I class the tokens of Ἀγαμέμνων with the adjectives, since it would seem to belong with the [mon-]stems in compounds that give rise to adjectives. A neuter $^{\text{X}}$μέμνα is unattested, however. Simply excluding the personal name from true membership in the category might be most appropriate.

[64] Likewise, $\mathscr{P}$ of [-mon-] adjectives without the inclusion of the tokens of Ἀγαμέμνων in that group would be $.02590674$.

[65] The high pragmatic potentiality of the [ma-]stems is perhaps indicative of the later substantial growth in types among this category; Buck and Petersen (1949: 221) report over $3600$ types in Greek literature.

following Probert's model, may be worthwhile. Typewise, [mon-]stem nouns in Homer are evenly divided between forms with recessive accentuation (like δαίμων [daímɔːn]) and those with high tone on the suffix (like λειμών [leimɔ́ːn]): there are eight with each accentuation (though suffix high tone is more common if the two recessively accented personal names Πάμμων [pámːɔːn] and Ἐχέμμων [ekʰémːɔːn] are excluded). Comparing the $\mathscr{P}$ calculated for [mon-]stem nouns with and without the two hapax personal names Πάμμων [pámːɔːn] and Ἐχέμμων [ekʰémːɔːn], the variable accentuation in this group of forms is easier to understand if we consider the category while excluding those two personal names; accentual variability makes more sense for a less parsable category; a $\mathscr{P}$ of $.02325581$ would fit better with an accentually invariable category, such as the [tu-]stems or [si-]/[ti-]stems.

Just as with Probert's [ro-]stem data above, a purely linear model cannot accommodate token frequency as a significant predictor of accentuation,[66] but a quadratic model using `token frequency` and `token frequency`$^2$ as predictors does.[67] The accuracy of this model is $71.4\%$, correctly predicting the accentuation for $10/14$ of the [mon-]stem nouns.[68] However, what is interesting to note is that the U-shaped relation between token frequency and accentuation is precisely the *opposite* of what Probert normally found: among the [mon-]stem nouns, the highest token frequency (δαίμων [daímɔːn] 'spirit', $76\times$) and lowest token frequency (πνεύμων [pneúmɔːn] 'lungs', $1\times$, or ἀλήμων [alɛ́ːmɔːn] 'wanderer', $2\times$) have recessive accentuation, while forms with middling token frequency (e.g., ἡγεμών [hɛːgemɔ́ːn] 'leader $26\times$, δαιτυμών [daitumɔ́ːn] 'guest' $9\times$) have non-recessive accent. The U-shape is, effectively, inverted; moreover, the relation seen between token frequency and accentuation is just the opposite of what the psycholinguistically grounded interpretation of token frequency would suggest.[69] For the record, these results are troubling: they suggest either a problem with the relation between token frequency and accentuation proposed by Probert, or that Homer as a corpus cannot always be regarded as an adequate representation of the true composition of the category.

One would hope that, in this particular case, the question of parsability and storage is instead inadequately addressed by the measure $\mathscr{P}$ for the category and the token frequency of the individual lexemes. Most immediately, we can consider both the synchronic availability of a base, and the relative frequency of that base and its derivative. If the criterion of basehood is left fairly loose, then most masculine [mon]-stems attested in Homer have a plausible base in Greek; four items, δαίμων, ἄκμων, λειμών, and χειμών are wholly without bases. It is evident that the availability of a base will not alone be a sufficient predictor: two of the baseless items are recessively accented (δαίμων and ἄκμων) while two are not (λειμών and χειμών). Base-derivative relative frequency does not offer clear guidance either: except for those four forms without bases, and the relative frequency of δαιτυμών [daitumɔ́ːn] 'guest' to δαιτύς [daitús] 'food', the base is always more frequent than the [mon-]stem derivative.[70]

[66] $Pr(> |t|) = 0.916$.

[67] $Pr(> |t|) = 0.0713$ and $0.0429$, respectively.

[68] The three incorrect predictions are for three lower token frequency lexically accented forms. All recessive forms are correctly predicted.

[69] Note, however, that Probert considered items attested $0\times$ in her corpus to represent the lowest token frequency members of the category.

[70] For determining base frequency, I simply took the token frequency of the entire lemma, e.g., the token

| Form | GLOSS | $N$ | Accent | Base | Base $N$ |
|---|---|---|---|---|---|
| ἡγεμών [hɛːgemɔ́ːn] | leader | 26 | lexical | ἄγω | 298 |
| κηδεμών [kɛːdemɔ́ːn] | guardian (of the dead) | 2 | lexical | κήδω | 43 |
| ἥμων [hέːmɔːn] | thrower | 3 | recessive | ἵημι | 180 |
| ἀλήμων [alέːmɔːn] | wanderer | 2 | recessive | ἀλάομαι | 41 |
| δαίμων [daímɔːn] | spirit | 76 | recessive | – | – |
| ἄκμων [akmɔːn] | anvil;heaven | 4 | recessive | – | – |
| βητ-άρμων [-armɔːn] | dancer | 2 | recessive | ἀραρίσκω | 70 |
| πνεύμων [pneúmɔːn] | lung | 1 | recessive | πνέω | 17 |
| δαιτυμών [daitumɔ́ːn] | guest | 9 | lexical | δαιτύς | 1 |
| τελαμών [telamɔ́ːn] | broad strap | 20 | lexical | τλάω | 82 |
| θημών [tʰɛːmɔ́ːn] | heap | 1 | lexical | τίθημι | 379 |
| κευθμών [keutʰmɔ́ːn] | hiding place, hole | 3 | lexical | κεύθω | 18 |
| λειμών [leimɔ́ːn] | meadow | 16 | lexical | – | – |
| χειμών [kʰeimɔ́ːn] | winter | 5 | lexical | – | – |

Table 7.4: [mon-]Stem Nouns in Homer

Nevertheless, a model including the ratio of base frequency to relative frequency, and that ratio squared, in addition to the token frequencies of the items, has better predictive power, correctly predicting the accentuation for $12/14$ items.[71] Use of relative token frequencies may then be a productive direction for exploration, just as Hay's work would suggest.

### 7.4.3 Greek [lo-]Stems

As an explicit comparison to Probert (2006b: Ch. 9)'s assessment of nouns built with the suffix -λο- [-lo-], I here assess their situation in Homer alone. Recall from 7.3.1 above that Probert openly identified [lo-]stems as an unproductive category in Greek, and attributed their overwhelming recessive accentuation, even among very high token frequency lexemes such as ὅπλον [hoplon] 'tool', to that unproductivity. Risch (1974: 110–2) offers a much more comprehensive list of forms possibly including a suffix [-lo-] than does Probert, who considers just masculine and neuter stems in [-los]/[-lon] (thus excluding feminines in [-lɛː], e.g., κεφαλή [kepʰalέː] 'head' or Σεμέλη [semélɛː] 'personal name'). Most crucially, Probert largely limits her list of [lo-]stems to items that seem etymologically to contain a suffix /-lo/, excluding items that happen to terminate in a sequence [-lo-]. For instance, χόλος [kʰólos] 'bile', Probert excludes because the [-l] belongs etymologically to the root ($<$ */ĝʰolh$_3$-o-s/], though Risch lists it among [lo-]stems.

frequency of the lemma τίθημι [títʰɛːmi] 'put' in Homer (379) as the base frequency relative to θημών [tʰɛːmɔ́ːn] 'heap'. A more precise identification of specific members of the verbal paradigm as the base might be helpful.

[71] This model also has a lower Akaike Information score than the model with token frequencies alone (21.645 vs. 25.113), which suggests that the relative frequency predictors are indeed adding value, not overfitting the data.

Which items to include in the assessment of an affix's productivity presents a more acute problem than for the categories already discussed, in which either a synchronic or etymological base for nearly all forms can be identified. In Table 7.5, I therefore present calculations of productivity for three different groups: one, including all masculine and neuter nouns in [-lo-] listed in Risch 1974: 110–2, including personal and geographical names; two, excluding personal and geographical names and the most frequent of the etymologically unsegmentable forms, μῆλον [mɛ̂ːlon] 'sheep, goat; apple'; third, just those lexemes that Probert considers to contain the suffix [-lo-], as given in Probert 2006b: 222–3.

| Category | $V$ | $N$ | $n_1$ | $\mathscr{P}$ | $S$ | $\mathscr{I}$ |
|---|---|---|---|---|---|---|
| All [-lo-] | 57 | 651 | 18 | .02764977 | 104.5647 | 1.834468 |
| [-lo-] w/o PN, μῆλον | 44 | 512 | 9 | .01757812 | 66.95684 | 1.521746 |
| Probert Items | 19 | 224 | 2 | .008928571 | – | – |

Table 7.5: Productivity Statistics for Homeric [lo-]Stems

The differences in the calculation of productivity between these groups are quite stark: whereas the inclusion of personal and geographical names results in a $\mathscr{P}$ greater than any of the Greek nominal categories examined so far apart from the [tu-]stems, reducing the set to just those items considered by Probert suggests a category of low to no productivity.[72] Since the appurtenance of proper names such as Σίπυλος [sípulos] or Σκῶλος [skɔ̂ːlos] to a category of derived [lo-]stems seems genuinely doubtful, Probert's assessment that [lo-]stems are unproductive would appear to be justified. The extent to which [lo-]stems are less productive than the other categories that Probert examined ([ro-]stems, [no-]stems) still requires confirmation, but the minimal productivity (and thus poor parsability) of [lo-]stems is entirely compatible with a tendency towards recessive accentuation.[73]

The interesting question is then why some few [lo-]stems, such as βηλός [bɛːlós] 'threshold' or μοχλός [mokʰlós] 'bar, lever', should exhibit non-recessive accentuation at all. Under the hypothesis that the highest token frequency lexemes would be most likely to preserve a historical non-default accent as part of a holistic lexical entry, it is surprising that all the [lo-]stems with non-recessive accentuation are of middling token frequency (e.g., χηλός [kʰɛːlós] 'large chest' $9\times$ in Homer and in Probert's corpus), and that we find no suffix-accented forms among the higher frequency [lo-]stems (e.g., ὅμῑλος [hómiːlos] 'crowd', $90\times$ Homer, $153\times$ in Probert's corpus), is likewise surprising.[74] Indeed, the situation appears similar to that of

[72] Indeed, calculating $S$ and $\mathscr{I}$ for the last group is impossible because the frequency distribution of the set violates the assumptions about the underlying structure of the data (i.e., that it is genuinely Zipf-distributed) to perform accurate extrapolations.

[73] $46/54$ [lo-]stem nouns given in Probert 2006b: 222–3 have recessive accentuation, including all $18$ items that have a frequency of $0$ in Probert's corpus.

[74] Perhaps even more amazing is that the [lo-]stems with high tone on the suffix in Ancient Greek that persist into Modern Greek, as far as I can determine, still have the primary stress in the same position as the high tone. Modern Greek has ομφαλός [omfalós] 'navel', μυελός [mielós] 'marrow', χυλός [çilós] 'porridge', and μοχλός [moxlós] 'lever'. These appear to be entirely standard lexemes in Modern Greek (i.e., they are not learned forms with an archaizing word prosody). See entries in Pring 1982.

the [mon-]stems above, in which items with neither high nor low token frequency exhibit non-recessive accentuation, contrary to Probert's general hypothesis.

### 7.4.4 Summary: Greek Categories

For convenience here, I repeat in 7.6 the statistics for the categories discussed above, ordered by $\mathscr{P}$ from greatest to least.

| Category | $V$ | $N$ | $n_1$ | $\mathscr{P}$ | $S$ | $\mathscr{I}$ |
|---|---|---|---|---|---|---|
| [-tú-] | 19 | 197 | 10 | .05076142 | 43.99002 | 2.315264 |
| [-si-]/[-ti-] | 46 | 422 | 9 | .02132701 | 50.48234 | 1.097442 |
| [-ma-] | 66 | 1108 | 20 | .01805054 | 95.4988 | 1.446952 |
| [-mon-] adj. | 19 | 377 | 5 | .0132626 | 23.66278 | 1.245409 |
| [-mon-] noun w/o PN | 14 | 170 | 2 | .01176471 | – | – |
| [-lo-] (Probert Items) | 19 | 224 | 2 | .008928571 | – | – |

Table 7.6: $\mathscr{P}$-sorted Productivity Statistics of Greek Categories

The clearest conclusion to follow from the morphological categories considered here is that the categories with lowest $\mathscr{P}$ ([mon-] and [lo-]stem nouns) are also the categories that exhibit type variation in accentuation across members of the category. This result is in accord with the predictions offered at the outset of this section: categories with lower $\mathscr{P}$ are less parsable, and therefore their members are less parsable, leading to cases in which accentuation changes at the level of the individual lexeme, because the prosodic property of the derivational morpheme (insofar as an unproductive derivational morpheme is learned at all) does not factor in to the computation of the word prosody. I therefore hold to the hypothesis that weak parsability of a category, as measured by $\mathscr{P}$, provides a necessary condition for accentual variability among members of a category.

At the same time, even this relatively limited examination of a few categories has revealed some further issues. Most importantly, token frequency as a predictor of accentuation assumes the opposite role for [mon-] and [lo-]stems as it did for Probert's [ro-]stems (discussed under 7.3 above). Recourse to the relations between bases and their derivatives did not offer an immediate account of the distribution of accentuation among [mon-]stem nouns, but the range of frequency relations that might impact parsability (especially morphological family size) remains to be explored. In addition, the degree of productivity measured for [ma-], [si-]/[ti-], and [tú-]stems in Homer appears to bear relatively little relation to the number of types that actually appear in those categories in the history of Greek. Finally, the consistently recessive accentuation of [si-]/[ti-]stems, which must be an innovation in Greek, does not appear to be attributable to wholescale "demorphologization" of that category's members. [lo-]stems may come close to instantiating the possibility of complete categorical demorphologization, but the persistence of non-recessively accented forms into Modern Greek (fn. 74) raises the question of why isolated and unparsable forms, without being terribly frequent, might maintain the same prosodic prominence across millennia.

### 7.4.5 Vedic *ti*- and *tu*-Stems

The Vedic categories cognate to those Greek noun-forming suffixes examined at 7.4.1 above present a more complex accentual situation, on the surface at least, than their Greek counterparts. Although *ti*-stems largely show the ictus in the derivational suffix, thus *-tí-* $(96/112$ types), a notable minority attest the ictus on the root $(16/112)$, and there exist some accentual doublets, e.g., both *tr̥ptí-* and *tŕpti-* 'satisfaction'.[75] *tu*-stems, which typically make masculine action nouns (e.g., *sótu-* 'pressing'), sometimes concretized (e.g., *sétu-* 'fetter'), exhibit considerable variability in accentuation; just under one-third $(10/33)$ show the ictus on the derivational suffix, while others are accented on the root.[76] The basic productivity statistics for the two categories are given in 7.7.

| Category | $V$ | $N$ | $n_1$ | $\mathscr{P}$ | $S$ | $\mathscr{I}$ |
|---|---|---|---|---|---|---|
| *-ti* | 112 | 2871 | 27 | .0094043891 | 161.3779 | 1.440874 |
| *-tu* | 33 | 733 | 8 | 0.01091405 | 41.43502 | 1.255607 |

Table 7.7: Productivity Statistics for RVic *ti*- and *tu*-Stems

In comparison to the baseline measure of a productive category given by sigmatic aorists in the RV ($\mathscr{P} = 0.05588822$), both of these categories appear to be unproductive. Indeed, the $\mathscr{P}$ and $\mathscr{I}$ for both *ti*- and *tu*-stems is below the values calculated for even *root* aorists ($\mathscr{P} = 0.01375095$, $\mathscr{I} = 2.426445$).

As I mentioned in fn. 41 above, /-ti-/ must be an unaccented suffix, because the ictus surfaces on *-mánt-* when in combination with that suffix, e.g., *r̥ṣṭí-* 'spear' → *r̥ṣṭimánt-* 'having a spear'. The same happens to be true for *tu*-stems as well, e.g., *yātú-* 'demon' → *yātumánt-* 'demonic'. The low productivity of these two derivational suffixes then suggests a straightforward solution to the variability of accentuation in these categories: *ti*- and *tu*-stems with ictus on the derivational suffix are morphologically parsed forms in which the presence of the suffix is recognized, and the ictus appears regularly on that suffix to satisfy HeadStress (e.g., /śru-ti-/ → *śrutí-*). On the other hand, *ti*- and *tu*-stems with root accentuation are unparsed forms in which no lexical accent is present, and thus the ictus in placed on the leftmost syllable (e.g., /śruti-/ → *śrúti-*).[77] As Lundquist (2015) notes, following earlier literature, *ti*-stems tend to adopt the ictus on the root during the history of Vedic, though the forms with

[75] I have counted otherwise identical forms with different accentuation as distinct types. I am very grateful to Jesse Lundquist for sharing with me the RVic frequency data on *ti*-stems that he collected.

[76] Infinitives such as *-tum* or *-tave* are not included among *tu*-stems here. I also exclude from the *tu*-stems two adjectives, *tapyatú-* 'hot, glowing', and *siṣāsátu-* 'desiring'. I include two feminines, *vástu-* 'morning, light' and *jívātu-* 'life', as well as one neuter, *vā́stu-* 'dwelling place'. The very frequent *krátu-* 'power' is probably best analyzed as an *u*-stem to a root */kret-/ in origin; without its tokens (281), the $\mathscr{P}$ of *tu*-stems would be substantially higher, at 0.01769912, though nevertheless below the productive zone of sigmatic aorists.

[77] *Pace* Lundquist 2014, I doubt that the frequency of *ti*-stems in compounds plays a substantial role in generating *ti*-stems with leftmost ictus. Although it is true on its face that the ratio of compound token frequency to simplex token frequency in the RV is much greater among *ti*-stem simplexes with initial-syllable ictus $(217/43$ vs. $827/1170)$, the token frequency of compounds corresponding to root-accented simplex *ti*-stems is very much skewed by a single outlier, *á-diti-* 'unboundedness = name of a goddess' $(174\times$ RV) alongside *díti-* 'bound-

the highest lemma token frequency in the RV (e.g., *ūtí-* 'aid' 289×, *dhití-* 'thought' 73×) tend to maintain the ictus on the suffix; I think that Lundquist is right to suggest that these forms with a long history of suffixal ictus should be regarded as preserving the non-default ictus on account of their high token frequency (i.e., the lexemes /ūtí-/ and /dhití-/ are learned and maintained). I have not seen evidence of any clearer trend towards initial-syllable ictus beyond the RV among the *tu*-stems, but given the lesser type and token frequency of that category, such evidence may not be forthcoming.

There are two outstanding questions that remain. First is whether the initial-syllable ictus among some members of these categories might be attributed to the lexical accent of a root. The existence of doublets among the *ti*-stems (such as *śaktí-* and *śákti-* 'skill, ability'), at least, excludes lexical accents on roots as a principal explanatory factor. Among the *tu-* stems, although I am not aware of any accentual doublets, we find forms with initial-syllable ictus to roots that show clear evidence of being unaccented: for instance, we find *hántu-* 'strike' alongside 3.pl.pres.act.ind. *ghnánti* (← /han-ánti/), and *gántu-* 'going, way' alongside 3.pl.aor.act.inj. *gmán* (← /gam-ánt/). The further issue concerns how the surface intial-syllable ictus is properly to be described. Namely, is the change from /śru-ti-/ > /śruti-/ (with no lexical accents), which then surfaces as [śrúti-], or from /śru-ti-/ > /śrúti-/ (with a lexical accent on the initial syllable), which likewise surfaces as [śrúti-]? One form possibly bears on this question: *śáktīvant-* 'strong', an adjective derived with /-vant-/ from *śákti-*.[78] Were *śáktīvant-* derived from /śakti-vant-/, we would expect to find [X]*śaktivánt-* (like *padvánt-* ← /pad-vant-/); instead the form implies that the UR is /śákti-vant-/ (like *gómant-* ← /gó-mant-/). Two adjectives in /-mant-/ derived from *tu*-stems also do not show ictus on *-mánt-*: *krátumant-* 'having power' and *pitúmant-*. These two forms imply /krátu-mant-/ and /pitú-mant-/, in constrast to /yā-tu-mant-/, which surfaces as *yātumánt-* (*yātú-* ← /yā-tu-/). *śáktīvant-*, for its part, suggests that "demorphologization" has created not merely an unaccented stem /śakti-/, but a lexically accented stem /śákti-/.

The reason for the selection of a UR /śákti-/ may find an explanation in the theory of **Lexicon Optimization** proposed by Prince and Smolensky (1993 [2002]: 209–14) (see also discussion, with references to related literature, in Krämer 2012: Ch. 8). Without engaging in detailed exposition here, I interpret Lexicon Optimization as a model of UR selection that selects the most harmonic UR from the set of URs compatible with the learning data. First, note that the constraints and weights used in the accentual grammar sketch under 7.2.3 above directly predict a paradigm without ictus alternation, given unaccented /śakti-/, as shown in the following examples.

(53) acc.sg. *śáktim* (2× RV)

---

edness' (3×). There is not a significant pattern of finding root-accented *ti*-stems where the *ti*-stem is more frequent in compounds. The abstraction of *ti*-stems from compounds, as Lundquist mentions, following Wackernagel and Debrunner 1954: 633, may be a source of some root-accented *ti*-stems.

[78] Why the *ī* is lengthened in this form is unclear; Arnold (1905: 127) claims that a short vowel, thus **śaktivant-* is to be restored, though Oldenberg (1909–12: 416) considers Arnold's explanation doubtful.

| śakti-m | | | MAX-IO(Accent) | HEADSTRESS$_{\text{Affix}}$ | IDENT-OO(Ictus, acc.sg.) | ALIGN-L(Ictus, PrWd) |
|---|---|---|---|---|---|---|
| a. | ☞ 0 | śáktim | | | | |
| b. | 1 | śaktím | | | | 1 |

(54) inst.pl. *śáktibhis* (2× RV)

| śakti-bhís | | | MAX-IO(Accent) | HEADSTRESS$_{\text{Affix}}$ | IDENT-OO(Ictus, acc.sg.) | ALIGN-L(Ictus, PrWd) |
|---|---|---|---|---|---|---|
| a. | ☞ 4 | śáktibhis | 4 | | | |
| b. | 6.5 | śaktíbhis | 4 | | 1.5 | 1 |
| c. | 5 | śaktibhís | | | $1.5 \times 2$ | $1 \times 2$ |

For the purposes of exposition at 7.2.3, I deliberately excluded the constraint DEP-IO(Accent), assuming it to be weighted lower than the other four constraints employed in the analysis so as to be irrelevant in actually deciding the position of the ictus. Nevertheless, it is clear that winning candidates such as *sáktim* or *śáktibhis*, just like acc.sg. *pā́dam* or *padvántam* (cf. examples (42) and (45) above) must violate DEP-IO(Accent). Hence, a learner confronted with *śáktim* and *śáktibhis*, perhaps generated just as in the two tableau immediately above, could be led to select /śákti-/ as the UR of the stem, because a stem /śákti-/ would not entail violations of DEP-IO(Accent). Thus, the winners *śáktim* and *śáktibhis* are rendered more harmonic by the acquisition of a UR /śákti-/ (and the losers, $^{\text{X}}$*śaktím* and *śaktibhís*, become correspondingly less harmonic, now subject to additional violations of MAX-IO(Accent)). Consequently, the derivation of an adjective /śákti-vant-/ (nom.pl. *śáktīvantas* 1× RV), results in leftmost ictus, just like, e.g., acc.sg. *gómantam* (cf. example (46) above).

(55) nom.pl. *śáktīvantas*

| śákti-vant-as | | | MAX-IO(Accent) | HEADSTRESS$_{\text{Affix}}$ | IDENT-OO(Ictus, acc.sg.) | ALIGN-L(Ictus, PrWd) |
|---|---|---|---|---|---|---|
| a. | ☞ 4 | śáktīvantas | | 4 | | |
| b. | 9 | śaktī́vantas | 4 | 4 | | 1 |
| c. | 6 | śaktīvántas | 4 | | | $1 \times 2$ |
| d. | 11 | śaktīvantás | 4 | 4 | | $1 \times 3$ |

### 7.4.6 Vedic *man*-Stems

Just as in Greek, Vedic possesses two varieties of suffix *-man-*, which are largely distinct in gender and accentuation. Vedic neuters in *-man-* build action and concrete nouns (e.g., *pátman-* 'flight', *bhū́man-* 'earth'), and seem to show the ictus on the root syllable exceptionlessly;[79] in accentuation, these neuters would seem to parallel the Greek neuters in -μα- [-ma-], which are exceptionlessly recessive. Masculines in *-mán*, meanwhile, build abstract and agent nouns (e.g., *bhūmán-* 'fullness', *dharmán-* 'orderer'), and show the ictus on the derivational suffix, with at least two exceptions in the RV, *óman-* 'companion' (so Grassmann 1872 [1976]) and *áśman-* 'stone'; this group parallels the Greek masculines in -μων [-mɔ:n], which sometimes show the ictus on the derivational suffix.[80] In Vedic, one also encounters pairs of a neuter in *-man-* alongside a masculine in *-mán-*, e.g., neut. *bráhman-* 'sacred formulation' : masc. *brahmán-* 'priest'.[81] The RV presents some eight such pairs, though in two such cases, neut. *várṣman-* : masc. *varṣmán-* both 'height' and neut. *svā́dman-* : masc. *svādmán-* both 'sweetness', no semantic difference is apparent; in the former case, neuter gender is uncertain, while the latter attests in the RV both masc. nom.sg. *svā́dmā* and neut. nom./acc.sg. *svā́dma*.[82]

Although collecting complete data on *man*-stems with the usual resources is no obstacle, I have followed Wackernagel and Debrunner 1954, rather than Grassmann 1872 [1976], in assigning the gender of numerous forms.[83] The fact of the matter is that determining the gender of many forms is not possible, in case the nom. or acc. happens to be unattested; in these cases the only guides are the semantics of the form and the accentuation itself. Consequently, the connection between accentuation and gender may have been less perfect than the lexicographers have assumed, but I see no real alternative. The usual statistics for neuter *-man-* and masculine *-mán-* appear in table 7.8.

The $\mathscr{P}$ found for neuters exceeds the $\mathscr{P}$ for *tu*-stems by just $\sim 8.5\%$, and is likewise less than the $\mathscr{P}$ calculated for root aorists. The masculine *mán*-stems, meanwhile, appear to be more productive than the respective neuters, though not productive at the level of even

[79] Grassmann (1872 [1976]) indicates a handful of forms with ictus on the suffix *-mán* as being neuters, namely, *premán-* 'love', *hemán-* 'zeal', *vidmán-* 'wisdom', and *prathimán-* 'breadth'. In fact, because all of these forms are (in the RV, at least) attested only in the inst.sg. or dat.sg., their gender is not self-evident. Wackernagel and Debrunner (1954: 754), meanwhile, identify all of these forms as masculines and (759) call the classification of *vidmán-* as a neuter "groundless."

[80] *óman-* occurs just $1\times$, in the inst.pl. *ómabhiḥ*, at RV 5.43.13b; Jamison and Brereton (2014) translate the form as 'succors', in which case it could easily be a neuter, with the expected accentuation of a neuter *man*-stem.

[81] Some lines of research describe such pairs as reflecting a process of referred to as "internal derivation", whereby a semantically related form is derived from a given base by a change of accentuation and ablaut pattern (see Widmer 2004: 30–5 for a brief overview of some types of internal derivation that have been proposed to have existed in PIE). With respect to the case at hand, one common view is that the masculines in */-mon-/ derive from neuters in /-men-/, thus (proterokinetic) */b$^{\text{h}}$lé$\hat{\text{g}}^{\text{h}}$men-/ (> *bráhman-*) → (amphikinetic ablaut) /b$^{\text{h}}$le$\hat{\text{g}}^{\text{h}}$món-/ (> *brahmán-*).

[82] In the case of *svā́dmā*, Oldenberg (1909–12: 79) considers possible either a lengthening of the *-ā* in a neuter *svā́dma**, or the existence of a masculine *svā́dman-*.

[83] Also following Wackernagel and Debrunner 1954, I have classed all stems in *-īman* as neuters, though the direct evidence for their gender is likewise almost entirely lacking.

| Category | $V$ | $N$ | $n_1$ | $\mathscr{P}$ | $S$ | $\mathscr{I}$ |
|---|---|---|---|---|---|---|
| Neuter *-man-* | 59 | 1267 | 15 | 0.01183899 | 80.58927 | 1.365924 |
| Masculine *-mán-* | 35 | 535 | 11 | 0.02056075 | 52.84702 | 1.509915 |

Table 7.8: Productivity Statistics for RVic *man*-Stems

*s*-aorists, for which $\mathscr{P} = 0.02953586$. These low levels of productivity accord with the fact that most *man*-stems, masculine and neuter alike, fall out of use by the Classical period (cf. Wackernagel and Debrunner 1954: 754–9).

Since Greek neuters in [-ma-] point to an unaccentable suffix /-ma-/, the same may be presumed for Vedic neuter /-man-/. Hence, even in case of loss of morphological structure and assignment of the ictus in accord with ALIGN-L(Ictus, PrWd), the surface result would be the same as for an parsed unaccentable suffix: /sad-man-/ and /sadman-/ would both surface as *sádman-* 'seat'. Moreover, the fact that we find no good evidence of a neuter *man*-stem with ictus on the derivational suffix may offer support to the hypothesis that, in Vedic, ictus at the right edge of the stem does not tend to develop among morphological classes that historically exhibited a left-edge ictus, whereas the opposite does occur (as in *ti-* and *tu*-stems above).

Conversely, while the RVic evidence is scant, we indeed find traces of initial syllable ictus among the masculine *mán*-stems: *óman-* 'companion' (1×; exists alongside *omán-* 'friendliness, aid'; but cf. fn. 80 above), *áśman-* 'stone' (29×), and possibly *várṣman-* 'height' and *svā́dman-* 'sweetness', if those two forms are not (solely) neuters, as mentioned above. The existence of some few masculines with initial syllable ictus accords with the hypothesis that the derivational suffix masc. /-man-/ is of high enough productivity to be largely parsable, but not so great as to be universally parsable, thus allowing for a few "demorphologized" forms with leftmost ictus.[84]

### 7.4.7 Vedic *iṣ*-Stems and Other Marginal Suffixes

As a final realm of exploration, I consider the status of three groups of lexemes with low type frequency. The possible suffixes *-(a)j* and *-it* are sufficiently rare that they might not be easily recognizable as such, if not for the fact that they contain forms with recognizable roots (e.g.,

[84]That the initial syllable ictus of *áśman-* 'stone', despite its agreement with Gk. ἄκμων [ákmɔ:n] 'heaven', is not the ictus of this form in PIE is shown by the position of the ictus in Lithuanian nom.sg. *ašmuõ*/*akmuõ* 'stone', which can only reflect a PIE *[$h_2$a$\hat{k}$mó$\acute{n}$], with ictus on the derivational suffix. Under the reconstruction of an amphikinetic paradigm *[$h_2$á$\hat{k}$mon-]/*[$h_2$ə$\hat{k}$mn-$\acute{}$], the Lithuanian accentuation is inexplicable; no known change affecting prosody in Baltic could have created *ašmuõ*. On the other hand, the initial syllable ictus attested in Sanskrit and Greek is compatible with "demorphologization" and default ictus assignment, departing from an original *[$h_2$a$\hat{k}$món-] ← */$h_2$e$\hat{k}$-món-/. It is worth noting that initial-syllable ictus *ákmuo* is attested in two Old Lithuanian texts (cf. Derksen 2015: s.v. *akmuo*). The fundamental point is that we have no means of accounting for an ictus on the derivational suffix, but the proposals here can handle a shift from non-initial to initial ictus.

*sanáj-* 'old', containing */sen-/as in Lat. *senex* 'old; old man').[85] The six forms terminating in *-aj* in the RV are both adjectives (*tṛṣnáj-* 'thirsty', *-svapnaj-* 'sleepy', *sanáj-* 'old', *dhṛsáj-* 'bold') and nouns (*bhiṣáj-* 'healer', *-dhvaj-* 'flag'); the same applies to the seven stems in *-it* (e.g., f. *sarít-* 'river', adj. *tadít-* 'nearby').[86] The rather better attested group of lexemes terminating in *-iṣ* are all neuter substantives. The *-iṣ-* stems in some instances have been thought to reflect neuters derived in PIE with a suffix */-s/ (perhaps an ablaut variant of the much more common suffix /-as/, PIE */-os/) to laryngeal-final roots (e.g., *kravíṣ-* 'gore' < */kréu̯H-s/, thus Schindler 1975), though most Vedic *-iṣ* stems are built to roots that are not laryngeal-final.

Accentually, all three groups are very stable: all *it*-stems show the ictus like in *sarít-*, the four *aj*-stems that occur outside a compound show the ictus as in *sanáj-*, and all but two of the 14 *-iṣ* stems show persistent ictus as in *kravíṣ-*; the three exceptions are *jyótiṣ-* 'light, shine', *vyáthiṣ-* 'way', and *ā́miṣ-* 'raw flesh'. The suffix /-íṣ-/, at least, gives evidence compatible with an accented suffix: *havíṣmant-*, *barhíṣmant-*, *śocíṣmant-*, though all three of these *mant*-stems are built to very high token frequency bases (252×, 142×, and 94×, respectively). However, the groups differ radically in their levels of productivity as assessed by $\mathscr{P}$: the *iṣ*-stems show the lowest value yet measured for any morphological category in Vedic, while the *aj-* and *it*-stems would be among the highest.[87] See the statistics calculated in 7.9.

| Category | $V$ | $N$ | $n_1$ | $\mathscr{P}$ | $S$ | $\mathscr{I}$ |
|---|---|---|---|---|---|---|
| *-iṣ-* | 14 | 750 | 2 | 0.0026666671 | 17.18024 | 1.22716 |
| *-aj-* | 6 | 22 | 3 | 0.1363636 | 13.26518 | 2.210863 |
| *-it-* | 7 | 37 | 3 | 0.08108108 | 11.46813 | 1.638304 |

Table 7.9: Productivity Statistics for RVic *iṣ-*, *aj-* and *it*-Stems

If *-ít* and *-áj* are indeed thoroughly parsable elements, then the ictus found on the suffix element is no surprise, whether those derivational suffixes are underlyingly accented or not. On the other hand, to find non-default accentuation among the wholly unproductive *iṣ*-stems is a surprise.[88] Some *íṣ*-stems are very frequent, such as *havíṣ-* 'oblation', which accords with the notion that higher token frequency forms may better preserve lexical accents in full-form storage (thus /havís-/), but the persistence of suffixal ictus among less frequent forms

[85] I say "possible suffixes" because it is unclear whether we are dealing with a unitary morphological category, or merely some group of lexemes that happens to contain an identical sequence of segments at the right edge of the stem, as well as the same pattern of accentuation. Wackernagel and Debrunner (1954) regard *it*-stems as a unitary group, but are skeptical of *-aj*.

[86] The complete list is: *tadít-* 'nearby' (2×), *dakṣinít-* 'with the right hand' (1×), *sarít-* 'river' (2×), *harít-* 'bay-colored (horse)' (26×), *divít-* 'shine, gleam (?)' (1×), *yoṣít-* 'young woman' (1×), *rohít-* 'red (horse)'.

[87] Given the very small number of types instantiated by *aj-* and *it*-stems, one must then accept that strong restrictions on the acceptable bases for those possible affixes were present, or that they had very limited pragmatic utility in our text (like Dutch *-erd*; cf. 2.2.2).

[88] Although one of the two forms without the ictus on the suffix, *jyótiṣ-* may have an accented root /dyáv-/ (cf. loc. sg. *dyávi* 'in heaven' ← /dyáv-í/, if /-íṣ-/ is accented as the *mant*-stems mentioned above might suggest, then [X]*jyotíṣ-* would be expected.

(e.g., *rocíṣ-* 'ray of light' $4\times$)[89] is then bothersome. I think that we must concede that, insofar as parsability can help to predict the accentuation of Vedic lexemes, the unproductivity of a derivational process instantiated in a lexeme is not a sufficient condition for unparsability.

### 7.4.8 Summary: Vedic Categories

Again for convenience, I repeat the statistics for the Vedic categories treated in the preceding sections, ordered by $\mathscr{P}$ from greatest to least, in 7.10.

| Category | $V$ | $N$ | $n_1$ | $\mathscr{P}$ | $S$ | $\mathscr{I}$ |
|---|---|---|---|---|---|---|
| *-aj-* | 6 | 22 | 3 | 0.1363636 | 13.26518 | 2.210863 |
| *-it-* | 7 | 37 | 3 | 0.08108108 | 11.46813 | 1.638304 |
| Masculine *-mán-* | 35 | 535 | 11 | 0.02056075 | 52.84702 | 1.509915 |
| Neuter *-man-* | 59 | 1267 | 15 | 0.01183899 | 80.58927 | 1.365924 |
| *-tu-* | 33 | 733 | 8 | 0.01091405 | 41.43502 | 1.255607 |
| *-ti-* | 112 | 2871 | 27 | 0.0094043891 | 161.3779 | 1.440874 |
| *-iṣ-* | 14 | 750 | 2 | 0.0026666671 | 17.18024 | 1.22716 |

Table 7.10: $\mathscr{P}$-sorted Productivity Statistics of Vedic Categories

We have seen that all of these categories, apart from the marginal *-aj-* and *-it-*, fail to exhibit an appreciable degree of productivity. Likewise, apart from those two marginal groups and the neuter *man*-stems, the other four groups show greater or lesser degrees of type variation in accentuation. In all four of those categories (*ti-*, *tu-*, *iṣ-*, and masc. *mán*-stems), to find the ictus on the derivational suffix is more common, but some forms with the ictus on the initial syllable could be found as well. In principle, given an accented root and an unaccented derivational suffix, to find the ictus on the root would be predicted; however, the existence of accentual doublets among *ti-* and masc. *mán*-stems, and the fact that some forms with initial-syllable ictus among the *ti-* and *tu*-stems contain roots that are demonstrably unaccented, rules out this possibility as a general explanation.

Instead, I have suggested that the forms among these categories with initial-syllable ictus are best explained as an effect of "demorphologization". The general plausibility of this account follows from the fact that among all these categories $\mathscr{P}$ is not very high (in relative terms for the RV); the implication is that the derivational suffix itself is correspondingly less parsable. Among the four categories with accentual type variation (*-ti-*, *-tu-*, *-mán-*, *-iṣ-*), there is a weak negative correlation between $\mathscr{P}$ and the proportion of types with initial-syllable ictus ($r = -0.4609134$).[90] Yet it is clear that low $\mathscr{P}$ does not automatically entail the adoption of default word prosody among all forms of a category. At present, the data

[89] *rocíṣ-* in fact occurs only $1\times$ in the RV as a simplex with the accentuation given here, but two instances of the compound *svarocíṣ-* confirm the accent.

[90] The correlation is much stronger ($r = -0.9271073$) if *tu*-stems are excluded; *tu*-stems have the highest proportion of types with initial syllable ictus $10/33$, but a higher $\mathscr{P}$ than *ti-* or *iṣ*-stems, which have lower proportions of types with initial syllable ictus ($16/112$ and $3/14$).

point to the view that poor parsability as measured by $\mathscr{P}$ is a *necessary* condition for "demorphologization" and prosodic regularization, but not a *sufficient* condition: if a derivational suffix itself is highly parsable, "demorphologization" should not be possible, but in the event of low parsability of a derivational suffix, an individual lexeme may nevertheless remain parsable, depending upon the full host of morphological and frequency relations in which it participates.

## 7.5 Future Directions

In the course of this chapter, beyond offering some demonstration of the practical issues involved in the analysis of Vedic and Ancient Greek prosody, I have argued for two general claims: under 7.2, I advanced an interpretation of accentual dominance effects in Greek and Sanskrit as a consequence of morphological headedness, following Revithiadou (1999); under 7.3 and 7.4, I have considered the extent to which the Baayenian measures of productivity, especially $\mathscr{P}$, *qua* measures of morphological parsability (per Hay and Baayen 2003), can help to account for instances of "demorphologization" and prosodic regularization, as proposed by Probert (2006b). I believe that the further data collected on nominal categories in Homer and the RV under 7.4 broadly supports the notion that low productivity means low parsability, and that low parsability can impact the realization of prosody, insofar as the prosody is dependent upon morphological structure. However, the fact that massive categorical changes in prosody do not appear to occur in cases of low parsability, but perhaps gradually, as documented for Vedic *ti*-stems, suggests that weak parsability of the derivational suffix itself may be a necessary, but not sufficient, condition that underlies prosodic changes at the level of individual lexemes.

The myriad problems discussed in this chapter, are, moreover, far from fully adequate characterizations of how prosodic prominence at the word level is computed in Greek and Sanskrit. Insofar as the goal is to construct a model that can predict the accentuation of a given lexeme with a high degree of accuracy, substantially more data on morpological relations and word frequencies must be systematically collected; I believe that Homer and the RV can serve as closed corpora toward this end. The outlines of the analysis offered under 7.2 also require further empirical testing, and much more work on accentuation in compounds is a desideratum. With respect to the effects of "demorphologization" in particular, it remains to be established whether the ultimate effect is to create *unaccented* stems or to create *accented* stems that accord happily with the phonologically preferred ictus (as decided by ALIGN-L(Ictus, PrWd) in Sanskrit and RECESSIVE ACCENT in Greek). I hope that the contents of this chapter prove of use to other scholars interested in pursuing these questions.

# CHAPTER 8

## Vedic Perfect Weak Stems of the Form $C_1eC_2$-

This chapter treats the productivity and consequent analogical extension of an inflectional subclass of verbs in Vedic Sanskrit, namely, perfects whose weak stem exhibits a form $C_1eC_2$- (i.e., vowel [Ce:C-]) and apparent absence of the reduplication that usually morphologically characterizes the Vedic perfect.[1] This inflectional subclass could be defined as productive, in a loose sense, insofar as it clearly expands within the attested history of Vedic at the expense of weak stems that transparently exhibit reduplication, of the form $C_1aC_1C_2$-. In distinction to preceding chapters, the focus here lies less with establishing the degree of productivity that this subclass exhibits than with uncovering the conditions that held in early Vedic, with as much precision as possible, that permitted the category's expansion. Indeed, as I will discuss in **8.2**, the corpus-based measures of productivity that served us well in treating a category that belongs to a distinct derivational category, are less readily useful in this case, because defining class membership depends almost exclusively on phonological conditions, rather than the presence of a given morpheme or the operation of a morphological process.[2] Instead, this chapter will operate with learning simulations (use of the Minimal Generalization Learner in particular) as the principal tool of investigation; my objective is to lay out a clear diachronic path, traceable point-by-point, that properly captures the membership of all forms known to belong to this subclass, without overgeneration.

The essential question is: how and why is a particular pattern of inflection able to expand its domain of application? What conditions must hold between already existing instantiators of a pattern and potential targets in order for the potential targets to undergo change? Earlier perspectives on the present problem, in particular Bartholomae 1885, seem to take for granted that the mere existence of a pattern, stated in its most general form, is a sufficient condition for the expansion of such patterns via morphological analogy. A relatively unconstrained model of such changes, however, deprives both the historical linguist and linguists generally of more interesting answers to more precise questions, in particular, why an expanding pattern should go so far and no further, or what might permit expansion in the first place. An undefined model lacks satisfying answers to both questions: a pattern might have expanded to a greater or lesser degree, or might not have ever undergone any expansion.

In operating with the principle of minimal generalization in this chapter, I adopt a deliberately hard-line view, but which does make for explicitly testable claims and predictions. First, because pattern-induction is taken to be minimal, supraminimal rules are usually less

[1] Paradigmatically, the "strong" stem occurs in the active indicative singular and all persons of the subjunctive; the "weak" stem occurs elsewhere.

[2] At minimum, we are dealing with a morphological process that is highly phonologically conditioned.

reliable, in that they admit of more exceptions. Therefore, the analogical expansion of a process is the form-by-form extension of the scope of minimal rules that, while perhaps different in substance, produce similar surface results. Furthermore, minimal generalization sets up specific limits on how widely a process might expand, depending upon the properties of the originally instantiating members with respect to targets. We will see that, for original sets composed of members that are crucially similar, analogical expansion is unlikely, precisely because those members are best supported by a powerful minimal rule, rather than a feeble general rule.

Much of this chapter, then, will be concerned with the pursuit of an original set of Vedic $C_1eC_2$- stems that can predict the extent of forms actually attested in Sanskrit. Novel here is the sequential use of the MGL to model successive generations of learning. In this way, a "cascading" effect of class membership can be obtained, as new, gradually more reliable and more general rules that support the $C_1eC_2$- pattern can be learned. Barring other substantial changes to the phonology or lexicon of a language, minimal generalization makes certain outcomes inevitable, given a particular starting state.

## 8.1 Formal Preliminaries and Attestation of $C_1eC_2$- Forms

This section briefly reviews the salient formal characteristics of the Vedic perfect and then introduces the basic data on $C_1eC_2$- forms, between the RV and Pāṇini. Since the syntax and semantics of the perfect are not of much relevance to the problem at hand, I simply refer the reader to recent discussion in Dahl 2010: Ch. 5 and Kümmel 2000: 65–94; Kümmel's work may also be consulted for more detailed treatment of specific perfect forms as well.

### 8.1.1 The Vedic Perfect: Form

One of the three principal derived tense-aspect stems of Vedic, alongside the present and aorist, perfect stems are built through reduplication, and inflected forms furthermore employ a set of inflectional endings entirely distinct from the set employed with present and aorist stems.[3] See Macdonell 1916 [1993]: 146–58 for paradigms and the details of various idiosyncrasies not discussed here; see Steriade 1988b for an attempt to formalize the major aspects of reduplication in the perfect.

In general, the reduplicant adheres to a monosyllabic template of the form CV, in which the C usually corresponds to the leftmost consonant of the base, and the V corresponds to the zero-grade vocalism of the base (as diagnosed by the vocalism seen in the PPP derived with the suffix /-tá-/). Thus to $\sqrt{svap}$ 'sleep', we find 3.sg.perf.act.ind. *suṣvā́pa* (PPP *suptá-*). Four systematic and principled exceptions to both the consonantism and vocalism of the reduplicant hold:

[3] Reduplication is far from rare in the derivation of verbal forms in Vedic generally: four other distinct verbal categories besides the perfect also exhibit reduplication as a means of stem formation. See Kulikov 2005, and literature cited there (to which I add Sandell 2011b on reduplicated presents) for a convenient overview.

- in case the root shows a falling sonority onset, e.g., $\sqrt{stu}$ 'praise', the lower sonority consonant appears in the reduplicant, thus 3.sg. *tuṣṭāva*, not [X]*suṣṭāva*.

- in case the reduplicating consonant is a velar (*k*(*h*), *g*(*h*), or *h*), the reduplicant shows instead the corresponding palatal (*c*, *j*, or *j*); thus, to $\sqrt{k\mathring{r}}$ 'make', 3.sg. *cakā́ra*, not [X]*kakā́ra*.[4]

- in case the initial consonant of the root is an aspirate, the reduplicant shows the corresponding non-aspirated consonant (Grassmann's Law); thus, to $\sqrt{bhaj}$ 'divide', the 3.sg. is *babhā́ja*, not [X]*bhabhā́ja*. This phenomenon, however, is not particular to the perfect, or even reduplicated formations alone in Vedic, but rather reflects a post-lexical constraint against aspirated consonants in adjacent syllables.

- roots whose zero-grade vocalism is a syllabic liquid ($\mathring{r}, \mathring{l}$) show the vowel *a* in the reduplicant; cf. again *cakā́ra* (PPP *kṛtá-*).

Note also that a number of perfect forms exhibit a long vowel in the reduplicant, e.g. 3.sg. *jāgāra* 'is awake' (cf. Gk. 3.sg.perf. ἐγρήγορε [egrɛ:gore], with "Attic" reduplication, but Av. 3.sg.perf. *jaγāra*, with short vowel in the reduplicant). General consensus holds that such long-vowel reduplicants in the perfect orginate to roots in an initial laryngeal (thus *jāgāra* $<$ PIE Transponat *[geH.gó.re]), but a compensentory lengthening historically accounts for only a small number of relevant forms, and a coherent and satisfactory account of long-vowel perfects in Vedic and Avestan is still a desideratum (despite the attempt of Krisch 1996).

The most relevant characteristic of the perfect for the problem to be treated here is its ablaut. While the vocalism of the reduplicant is fixed, the base undergoes vowel alternations: in principle, the strong stem takes *guṇa* (full grade, presence of a vowel *a* or *ā*) and the weak stem takes *loka* (zero grade, absence of a vowel *a* or *ā*). The word ictus also exhibits alternation, falling on the base in the strong stem, but to the right of the base, on an inflectional ending or the optative morpheme *-yā/-ī-*, in the weak stem. All of the 3.sg.act. forms cited above (with ending /-a/) instantiate the strong stem; corresponding weak stems, given here in the 3.pl.act., would be: *suṣupúr*, *tuṣṭuvúr*, *cakrúr*, *jugupúr*, and *bhejúr*.

### 8.1.2 Attestation of $C_1eC_2$- Forms

The 3.pl.perf *bhejúr* presents us with our first example of a $C_1eC_2$- form: a root of the shape $C_1aC_2$- ($\sqrt{bhaj}$), in its weak stem, takes this shape, without evident reduplication. Expected instead, both from the historical point of view and according to the general synchronic process of weak stem formation in Vedic, would be a stem of the form $C_1aC_1C_2$-, in this case, [X]*babhj-*. Forms such as *bhej-* occur, already in the RV, to some fifteen distinct roots; those stems and their token frequencies are given in Table 8.1.[5]

Six further types are first attested in the *Atharvaveda*; eight others appear first in later Vedic texts. Finally, another eight perfect weak stems of this sort come to light in Epic or

[4] Although this alternation between velar and palatal originates in palatalization before [+front] vowels (e.g., cakā́ra $<$ PIE *[kekóre]), in Vedic it is a general constraint on the form of the reduplicant, as reduplicants

| ROOT | Gloss | Perfect Weak Stem | RV Stem Token Frequency |
|---|---|---|---|
| $\sqrt{tap}$ | 'be warm; heat' | *tep-* | 2 |
| $\sqrt{dabh}$ | 'deceive' | *debh-* | 2 |
| $\sqrt{nam}$ | 'bow' | *nem-* | 1 |
| $\sqrt{pac}$ | 'cook' | *pec-* | 1 |
| $\sqrt{pat}$ | 'fly, fall' | *pet-* | 3 |
| $\sqrt{bhaj}$ | 'divide, share' | *bhej-* | 7 |
| $\sqrt{yaj}$ | 'worship' | *yej-* | 3 |
| $\sqrt{yat}$ | 'take a position' | *yet-* | 5 |
| $\sqrt{yam}$ | 'stretch out; hold' | *yem-* | 34 |
| $\sqrt{rabh/labh}$ | 'grasp' | *r/lebh-* | 3 |
| $\sqrt{śak}$ | 'create, shape' | *śek-* | 4 |
| $\sqrt{śap}$ | 'curse' | *śep-* | 2 |
| $\sqrt{sad}$ | 'sit' | *sed-* | 31 |
| $\sqrt{sap}$ | 'care for, honor' | *sep-* | 1 |
| $\sqrt{sah}$ | 'overpower, win' | *seh-* | 3 |
| **TOTAL** | – | 15 | 102 |

Table 8.1: Perfect Weak Stems in $C_1eC_2$-in the *Ṛ̊gVeda*

Classical Sanskrit. These forms are all listed, with the earliest attestation of a $C_1eC_2$- form, in Table 8.2. Note that the first attestation of a $C_1eC_2$- form merely establishes a *terminus ante quem* for the creation of that form; in some cases, the weak stem, or indeed any form of the perfect, is unattested before that time. To speak of the "first attestation" for a $C_1eC_2$- weak stem thus in no way necessarily implies the prior existence of a weak stem of the form $C_1aC_1C_2$-. Except for the RV, the AV in its Śaunaka recension, and the epics *Mahābhārata* and *Rāmāyaṇa*, token frequency data is unavailable for these forms, but is in any case not crucial for present purposes.[6]

Indeed, the selection of precisely which roots are subject to taking a stem form $C_1eC_2$- is sufficiently complex that Pāṇini devotes seven sūtras to the problem. The principal sūtra is 6.4.120 (trans. von Böhtlingk 1887 [1964]):

with *u* vocalism demonstrate, e.g., 3.sg.perf. *jugopa* to $\sqrt{gup}$ 'protect'.

[5] Even in this tiny sample size, properties of usual lexical frequency distributions emerge: a low median, the mode is 1 (tied with 3), and a relatively large standard deviation (10.57) relative to the sample size (15 types, 102 tokens). Perhaps a more adequate way of characterizing this distribution, however, would simply be to say that it contains two large outliers (viz., the token frequencies of *sed-* and *yem-*).

[6] Pāṇini also cites another nine roots that at least "optionally", in his formulation, can take a perfect weak stem with *e* vocalism. Perfect weak stems with *e* vocalism are not attested to these roots in Sanskrit outside of Pāṇini's citations. Five of those nine roots are roots of the form $\sqrt{CRaC}$, which, like the independently attested stems *śrem-* and *bhrem-*, I do not propose to account for here. The four others, *jer-* (to $\sqrt{jar^{i}}$ 'grow old'), *phel-* (to $\sqrt{phal}$ 'burst, fructify'), *pheṇ-* (to $\sqrt{phaṇ}$ 'go'), and *redh-* (to $\sqrt{rādh}$ 'harm'), are predicted to assume $C_1eC_2$- weak stems under the final model described under 8.5.2. Perfects to $\sqrt{jar^{i}}$ and $\sqrt{rādh}$ are in fact included throughout in the models used here, because they attest perfects in Vedic Sanskrit.

| ROOT | Gloss | Perfect Weak Stem | First Attestation of $C_1eC_2$ Form |
|---|---|---|---|
| $\sqrt{car^{i}}$ | 'move' | *cer-* | AV |
| $\sqrt{tan}$ | 'stretch' | *ten-* | AV |
| $\sqrt{bandh}$ | 'bind' | *bedh-* | AV |
| $\sqrt{math^{i}}$ | 'rob' | *meth-* | AV |
| $\sqrt{man}$ | 'think' | *men-* | AVP |
| $\sqrt{sac}$ | 'follow' | *sec-* | AV |
| $\sqrt{naś}$ | 'perish' | *neś-* | ŚB |
| $\sqrt{nah}$ | 'bind' | *neh-* | AB |
| $\sqrt{pad}$ | 'fall, go' | *ped-* | AB |
| $\sqrt{mad}$ | 'rejoice' | *med-* | JB |
| $\sqrt{ram}$ | 'be content' | *rem-* | BĀU |
| $\sqrt{śram}$ | 'make/become tired' | *śrem-* | ŚB |
| $\sqrt{cam^{(i)}}$ | 'sip' | *cem-*[a] | E |
| $\sqrt{jap}$ | 'whisper' | *jep-* | E |
| $\sqrt{tar^{i}}$ | 'cross over' | *ter-* | E |
| $\sqrt{nad}$ | 'sound' | *ned-* | E |
| $\sqrt{ran}$ | 'ring' | *reṇ-* | E |
| $?\sqrt{ras}$ | 'roar' | *res-* | E[b] |
| $\sqrt{rāj}$ | 'be kingly; shine' | *rej-* | E |
| $\sqrt{lap}$ | 'prattle' | *lep-* | E |
| $\sqrt{dah}$ | 'burn; extinguish' | *deh-* | C.Skt. |
| $\sqrt{bhram}$ | 'wander' | *bhrem-* | C.Skt. |
| $\sqrt{laṣ}$ | 'desire' | *leṣ-* | C.Skt. |

[a] Possibly attested only once, MBh 5.81.59c; the Poona critical edition (Sukhtankar et al. 1971–1975) prints *ācemuḥ*, but the manuscripts differ substantially, and a different reading cannot be excluded.

[b] Cited as occurring in Epic by Whitney (1885 [1963]), but I cannot independently confirm the attestation.

Table 8.2: Perfect Weak Stems in $C_1eC_2$- beyond the *Ṛgveda*

> "Für ein zwischen zwei einfachen Consonanten stehendes अ eines Verbalstammes wird vor den Personalendungen des Perfects, die ein stummes क् haben (s. 1, 2, 5), ए substituirt, wenn für den Anlaut der Wurzel in der Reduplication kein anderer Consonant substituirt wird; bei dieser Substitution fällt der Reduplication ab."[7]

Sūtras 6.4.121–5 enumerate specific roots that (optionally) show the "substitution" of *e* for *a*, though they do not belong to roots that have the shape $C_1aC_2$-; sūtra 6.4.126 further specifies that roots beginning with *v-* and that roots with vocalism other than *ă* in their *loka*

[7] "An ए (*e*) is substituted before the personal endings of the perfect that have a "mute क् (*k*)" (i.e., "weak" endings) for an अ (*a*) that stands between two simple consonants of a verbal stem, if no other consonant is substituted for the initial consonant of the root; in the case of such substitution (i.e., of *e* for *a*), the reduplication falls away."

(zero grade) do not build such $C_1eC_2$- weak stems. Setting aside the cases of $C_1r/leC_2$- (e.g., *śrem-*), which tend to show the substitution of *e* "optionally", Pāṇini's formulations are entirely adequate: the perfect weak stems of biconsonantal roots, by the period of Epic Sanskrit at latest, exceptionlessly show $C_1eC_2$-, except to *v-* initial roots (which instead usually show *saṃprasāraṇa*, e.g., *ūh-* to $\sqrt{vah}$ 'travel, bring') and roots with a velar∼palatal alternation (e.g., *cakr-* to $\sqrt{kr̥}$ 'make' or *jagm-* to $\sqrt{gam}$ 'come, go').

The existence of these $C_1eC_2$- weak stems would then appear to be a mere morphological peculiarity of Sanskrit (for which one might nevertheless seek a phonological explanation). However, in a few cases, an older perfect weak stem of the shape $C_1aC_1C_2$- is attested alongside a younger $C_1eC_2$- form. Such cases demonstrate that the domain of the $C_1eC_2$- pattern must undergo some sort of analogical extension within the history of Sanskrit:[8]

- To $\sqrt{pat}$ 'fly, fall': *papt-* (RV Books II, V, IX)[9] vs. *pet-* (RV Books I, VIII, X)[10]
- To $\sqrt{tan}$ 'stretch': *tatn-* (RV) vs. *ten-* (AV+).
- To $\sqrt{mad}$ 'rejoice': *mand-* (RV) vs. *med-* (JB).[11]
- To $\sqrt{man}$ 'think': *mamn-* (RV) vs. *men-* (AVP+).
- To $\sqrt{sac}$ 'follow': *saśc-* (RV) vs. *sec-* (AV+).
- To $\sqrt{tar^i}$ 'cross over': *titir-* and *tutur-* (RV) vs. *ter-* (MBh).

For all other roots of the shape $C_1aC_2$-, only a $C_1eC_2$- perfect weak stem is known, from its earliest attestation (as is the case for, e.g., RV *sed-*, *bhej-*, *pec-*).

The principal objective of this chapter, from this point, is to identify the underlying causal conditions in Vedic that must have held in order for forms that we know (by direct attestation) or justifiably believe (by reconstruction) to originally have exhibited perfect weak stems of the form $C_1aC_1C_2$- to have developed weak stems of the form $C_1eC_2$-. What must have been true for, e.g., *pet-* and *men-* to have replaced *papt-* and *mamn-*, respectively? In section 8.2, I will first explore reasons why corpus-based measures of morphological productivity are not very meaningful for the study of this particular question. Morphological learning simulations, on the other hand, provide precisely the necessary tool to understand the expansion

[8] The other possibility is that the phonological (and phonotactic, specifically) grammar of Vedic underwent changes between the very earliest Vedic (RV Family Books) and later phases. Namely, forms like **papc-* were rendered illicit by interactions between phonotactics and the phonology of reduplication already by the RV Family Books. The more faithful form *papt-* was allowed; a tightening of phonotactic markedness would eventually render an output **papt-* illicit and so produce *pet-* instead. An account of this sort is considered more fully under 8.5.3.

[9] That is, relatively earlier portions of the RV.

[10] That is, relatively later portions of the RV.

[11] Kümmel (2000: 356–7) judges the two instances of 3.pl. ((*abhí*)*pra*)*mandúr* (as well as the middle participle *mandāná-*, 17x) that occur in the RV as perfect weak stems to $\sqrt{mad}$; Lubotsky (1998: 1035), on the other hand, considers the forms to be root aorists built to the neo-root $\sqrt{mand}$ (itself abstracted from a nasal infix present to $\sqrt{mad}$). I follow Kümmel; the translations of Jamison and Brereton 2014 are generally compatible with reading as either aorist or perfect forms.

of an inflectional subtype (compare the study of Spanish diphthongizing verbs in Albright 2008b); the MINIMAL GENERALIZATION LEARNER will therefore be the principal tool in this study.

## 8.2 Measuring the Productivity of an Inflectional Subclass

The core frequency statistics for $C_1eC_2$- perfect weak stems in the RV are as follow in Table 8.3, with a calculation of $\mathscr{P}$:

| $N$ | $V$ | $n_1 \sim \mathscr{P}*$ | $\mathscr{P}$ |
|---|---|---|---|
| 102 | 15 | 3 | .0294 |

Table 8.3: Frequency Statistics for RV $C_1eC_2$- Forms

The number of hapax legomena that belong to this inflectional subcategory contribute very little to the overall number of hapax legomena in the RV (small $\mathscr{P}*$), and indeed, contribute relatively little to the overall vocabulary growth of the RV, with so few types. Given the category's small token frequency, however, $\mathscr{P}$ takes on a fairly large value, indicative of a productive word-formational type. However, it seems reasonable to state that, for Pāṇini, this word-formational type has exhausted its potential productivity — every form that ought to show a stem $C_1eC_2$- in fact does. Precisely because this word-formational type has a strict formal limitation (i.e., it applies only to roots of the shape $C_1aC_2$-), and its productivity is directly dependent upon the productivity of the perfect itself, it is simply axiomatic that a perfect weak stem built to a root $C_1aC_2$- take the form $C_1eC_2$-; hence, the rate of increase for $C_1eC_2$- weak stems should be a simple function of the productivity of the perfect itself and the rate at which new $C_1aC_2$- roots enter the language:

$$Productivity(C_1eC_2-) = Productivity(perfect(Productivity(C_1aC_2-))) \quad (8.1)$$

These sorts of structural limitations on productivity are intuitively of a different nature than restrictions on, e.g., the acceptable syntactic base for an affix, and are indeed part and parcel of traditional grammatical description. As part of a paradigmatic, rather than syntagmatic relation, we (and Pāṇini) "know" that a Classical Sanskrit perfect built to a root $\sqrt{C_1aC_2\text{-}}$ will have the form $C_1eC_2$-. In this sense, the paradigmatic (inflectional) productivity seems more fully "realized" than syntagmatic (derivational) productivity. In other words, inflectional paradigms (despite "gaps") may be psychologically more complete than derivational processes.

More to the point, however, the corpus-based productivity measures that provide a good sense of the productivity of derivational categories or kinds of inherent inflection (as opposed to contextual inflection; cf. Gaeta 2007: 183–7 and references there) are not very meaningful for inflectional subtypes.[12] The $C_1eC_2$- subclass of Vedic perfects with which we

[12] By *contextual* inflection, I mean inflection that is determined largely by surrounding syntactic or phono-

| Inflectional Ending | $N$ | $V$ | $n_1$ | $n_2$ | $\mathscr{P}$ |
|---|---|---|---|---|---|
| 1.sg.pres. *-mi* | 331 | 106 | 57 | 15 | 0.172205 |
| 1.pl.pres. *-masi* | 138 | 52 | 31 | 8 | 0.2246377 |
| 1.du.impf.mid *-vahi* | 2 | 2 | 1 | 1 | $.\bar{3}$ |
| 3.pl.impv.act. *-antu* | 423 | 100 | 50 | 19 | 0.1182033 |
| 2.sg.subj.mid. *-dhvai* | 3 | 1 | 0 | 0 | 0 |

Table 8.4: Frequency Statistics for Five Verbal Inflectional Endings in the RV

are concerned here is, in effect, contextually determined by the phonological shape of the verbal root involved. Thus the $\mathscr{P}$ rating for a contextually determined inflectional subtype may still be an indicator of that subtype's parsability, but tells us little about where and why the subtype might take in new members.

What we may conclude is that the corpus frequencies of inflectional forms are not directly indicative of the productivity of such contextually determined inflectional processes. The CONFIDENCE measure employed in Minimal Generalization Learning seems much more adequate to this task, because that measure directly reports on the successful applications of a morphological process. But given that the MGL does not rely on any token frequency information, are we effectively asking: is only type frequency relevant to such contextually determined inflectional productivity? This assumption encounters problems in the face of psycholinguistic evidence: high token frequency inflectional forms may more weakly instantiate the prototype of an inflectional process (since even inflected forms may obtain high token frequencies), and thereby contribute less to the psychological activation of a given process (following the assumptions of the Morphological Race Model; cf. 3.3). The psycholinguistic literature offers us no reason to think that inflectional and derivational forms, in either processing or production, are fundamentally different; both are subject to the same effects of token frequency (cf. Bertram et al. 2000).

Nevertheless, specific inflectional forms typically (though of course not always) exhibit low token frequencies – as expected, the high number of low token frequency inflected forms entails parsing of the inflectional morphemes, and thus the productivity (due to high psychological activation) of those inflectional morphemes. Indeed, inflectional endings, as expected, show enormous values for $\mathscr{P}$; Table 8.4 gives some statistics on some verbal inflectional endings of Vedic in the RV.[13]

Generally, these examples suggest an inverse correlation between usefulness ($\mathscr{U} = V$; cf. 2.2.2) and $\mathscr{P}$ (removing the results for *-dhvai*, which seems to have a low $\mathscr{P}$ by accident of attestation, $r^2 = 0.5198$): higher $\mathscr{P}$ corresponds to lower type frequency. This result

logical factors, e.g., the marking of number on verbs or case on nouns. By *inherent* inflection, I mean inflection that follows generally from semantic content of an utterence, e.g., number on nouns or tense on verbs.

[13] Statistics were obtained using a text file of the RV *padapāṭha*, broken on whitespaces, and then searched using regular expressions of the form `.*INFL$` where `INFL` stands for the sequence of characters comprising that particular ending. Results were inspected and corrected by hand to weed out hapax legomena produced by forms with preverbs.

has a sensible interpretation: precisely because endings such as 1.sg.pres.act.ind. *-mi* are more useful, they apply to a large number of types and create a larger number of tokens, thereby exhausting more of their potential productivity; conversely, less useful endings like *-vahi* have saturated less of their potential productivity. This situation recalls the case of the derivational suffixes *-er* and *-ster* that form agent nouns in Dutch (cf. Baayen 2009). *-ster* forms only female agent nouns, and is dispreferred to "unmarked" *-er* except where the fact that the agent is female is emphasized; consequently, although *-er* and *-ster* have identical domains of productivity, *-er* has used up much more of its potential productivity than *-ster*, and thus measures as less productive.

Since contextually determined inflectional productivity, at least as far as the measure $\mathscr{P}$ is concerned, perhaps cannot be separated from usefulness, treating inflectional productivity solely in terms of type frequency then arguably does make sense. The sort of competition between PDE *flung* and *flinged* is better examined from the reliability of a morphological process determined by type frequency (e.g., how many context-determined types instantiate the vowel change [ɪ] → [ʌ] versus Ø→ [əd], under specific phonological conditions) than by the measure $\mathscr{P}$. Thus the same principle applies to the Vedic perfect weak stems under discussion here.

## 8.3 Previous Analogical Accounts of $C_1eC_2$- Forms

This section summarizes two existing accounts concerning the history and development of $C_1eC_2$- perfect weak stems, and argues that both are demonstrably inadequate. Specifically, I will claim, on the basis of psycholinguistic research into the processing of morphologically complex words (as summarized and reviewed in Chapter 3), that Lubotsky (2013)'s approach is categorically untenable, while the older view proposed in Bartholomae 1885 cannot be sustained when precise modeling is attempted. Under 8.3.3, I describe the preparations for learning simulations and execute an initial simulation to test the efficacy of Bartholomae's account. The preliminary results obtained in this section already nicely illustrate the perils of positing analogical changes without a restrictive and predictive model of how an analogy ought to behave.

### 8.3.1 Bartholomae 1885

The traditional approach to the creation of $C_1eC_2$- perfect weak stems goes back to Bartholomae (1885), who uses discussion of the subclass as a springboard to discussion of the developments of Proto-Indo-Iranian voiced sibilants in Indic. The two issues are directly connected, since, Bartholomae judges, the stem *sed-* (to $\sqrt{sad}$ 'sit') that derives from PIIr. *[səzd-], by loss of *[z], is a crucial member of the perfect weak stems to originally show such *e* vocalism.[14] Once the Proto-Indo-Iranian diphthong *[əj] monophthongized to [eː] in Vedic, the

[14] Bartholomae assumes that the Avestan perfect stem *hazd-* directly continues PIIr. *[səzd-]. I will suggest under 8.6.1 that this view cannot be taken for granted: Av. *hazd-* can readily reflect the synchronically productive formation of a perfect weak stem to $\sqrt{had}$ in Avestan.

weak stems *[jəjt-] and *[jəjm-] became *yet-* and *yem-*, thereby creating a small subclass that could, in principle, analogically expand its scope. Bartholomae takes the three stems *sed-*, *yet-*, and *yem-* as the historical basis for the entire inflectional subclass; this standard doctrine is repeated in detail most recently in Kümmel 2000: 19.

Nevertheless, among all the roots that build a $C_1eC_2$- form, only six (see the list on pg. 216 above) actually ever attest a weak stem of the form $C_1aC_1C_2$- (e.g., *papt-* or *mamn-*); the process of analogical extension appears to largely precede the composition of the earliest portions of the RV, and be complete in the later RV or AV. That the analogy be prehistoric is not a problem in itself, but, at minimum, a precise explanation of this problem should be able to account for why the RV Family Books would exhibit some supposedly analogical forms (*pec-*, *bhej-*, *śek-*, *sep-*) alongside reduplicated forms (*papt-*, *mamn-*). If the texts indeed show that speakers of early Vedic simultaneously used stems like *sep-* alongside stems like *papt-*, one would ideally like to know why a **sasp-* would have been replaced by *sep-* earlier than *papt-* by *pet-*. While the set of perfect weak stems that exhibit a *lautgesetzlich e* vocalism under Bartholomae's account is small, it has the virtue of proceeding from a set of forms (*sed-*, *yet-*, *yem-*) from which an abstract pattern, of greater or lesser specificity, could plausibly be abstracted and applied to other forms. Thus, there is no *a priori* theoretical reason to reject Bartholomae's account — it must be tested empirically.

### 8.3.2 Lubotsky 2013

In a recent article, on the other hand, Lubotsky (2013) proposes that the base of analogical extension is limited to just a single form, *sed-*. More specifically, Lubotsky suggests that the high token frequency of this stem (*sed-* 31x RV) makes it analogically influential: "…the analogical spread of -e- in the weak grade of the perfect can only be attributed to *sasāda* / *sed-*. This seems unproblematic to me. The weak forms with -e- are very well attested for this common root…". This claim does not arise on its own, but is rather a logical consequence of the argument that Lubotsky's paper advances: that the PIIr. diphthongs *[əw] and *[əj] were not monophthongized to [oː] and [eː] at the time of the composition of the earliest Vedic texts.[15] Since Lubotsky accepts that the Vedic stems *yet-* and *yem-* are ultimately direct continuants of PIIr. *[jəjt-] and *[jəjm-] by sound change, yet forms like *pec-* appear prior to the period to which Lubotsky would date the monophthongization of diphthongs, Lubotsky is left with *sed-* as the sole source by which to generate forms like *pec-*. As a relative chronology, the developments for Lubotsky are as follows:

1. *[səzd-] > *[səːd-] (by loss of voiced sibilants, with compensatory lengthening).
2. Analogical extension of the vocalism *[əː] from *[səːd-] to other perfect weak stems (*[bəbhj́-], *[pəpč-], etc. >> *[bhəːj́-], *[pəːč-], etc.).
3. Monophthongization of *[əj] > [eː] (thus, *[jəjt-], *[jəjm-] > *yet-*, *yem-*).

[15] "…the poets still pronounced the diphthongs (**ai* and, by extension, **au*), so that the monophthongization must at least be posterior to the compilation of the AV." In fact, the idea that PIIr. *[əj] and *[əw] are still present in early Vedic appears also in 19th-century scholarship; cf. the citations in Wackernagel 1896: 39.

4. Merger of *[əː] with [eː] (thus *[səːd-], *[bhəːǰ-], *[pəːč-] > *sed-*, *bhej-*, *pec-*).

Unfortunately, Lubotsky's reasoning in this scenario surrounding *[səːd-] (*sed-*) runs precisely counter to the psycholinguistic evidence from morphological processing discussed in Chapter 3. To reprise briefly: a good analogy (i.e., a productive extension of a morphological process) is based on a substantial number of different *types*, many of which should exhibit *low*, not high, *token* frequency — the members of the class must be parsable as instantiating a morphological process in order for that process to be learnable. Under Lubotsky's account, *sed-* would simply be a morphologically isolated unique type — no actual pattern that could build perfect weak stems by replacing *a* with *e* in the root would have existed. Moreover, precisely because the stem *sed-* is so frequent, it, like high token frequency forms generally, is more likely to have been produced and processed from memory than have been generated or analyzed online; *sed-* would have a status closer to an independent lexeme. Under the Morphological Race Model proposed in Frauenfelder and Schreuder 1992 (followed by Hay and Baayen 2002), the correlation of high token frequency with low productivity is precisely a consequence of the fact that forms generated or processed from memory contribute little to the representation of a more abstract and extensible process. *sed-* alone could in no way directly beget further perfect weak stems with *e* vocalism.

### 8.3.3 Evaluating the Traditional Account with the MGL

The fact that unique types make poor, if not impossible, bases for an analogy is also built into the behavior of the Minimal Generalization Learner: the learner will not apply morphophonological rules posited on the basis of fewer than two types. Such unique rules are simply coextensive with the context of the single type that instantiates the rule; such unique types are liable to replacement by higher confidence rules, and thus must be presumed to persist only because they are lexicalized (accessed from memory), or reflect the operation of further purely phonological processes (i.e., *sed-* might conceivably derive from morphologically derived /sa-sd-/, because phonological constraints would exclude *[sa-zd-] or *[sasd-]). Hence, a learning simulation that contains only *sed-* as an exemplar for $C_1eC_2$- forms would not predict any analogical extension whatsoever. Conversely, Bartholomae's set of three stems (*sed-*, *yet-*, *yem-*) does contain the potential for the abstraction of a rule that might apply to further forms. The objective of this section is therefore to evaluate the predictions and performance for a history built on those three stems; I will conclude that such a basis is grossly inadequate.

#### 8.3.3.1 Preparing the MGL Simulations

**Determining the Base**

Various tests of the MGL (see generally the literature cited in 2.3) have examined two distinct types of analogies: intraparadigmatic and interparadigmatic. Intraparadigmatic analogies show the extension of a surface stem form present within one paradigm slot to other forms within the paradigm; the Latin *honor* analogy (cf. Albright 2002b: Ch. 3), or the spread

of the stem form of the 1.sg. in Yiddish present tense inflection (cf. Albright 2010) clearly constitute intraparadigmatic analogies. Intraparadigmatic analogies depend crucially upon the determination of a given synchronic base. Interparadigmatic analogies, on the other hand, result from the extension of a pattern mapping seen in the paradigms of some lexical items to other lexical items; these are analogies that depend upon the increased scope and/or greater reliability of the rules themselves. The limited productive extension of vowel alternation between present and preterite in PDE (cf. Albright and Hayes 2003), and the extension of diphthongization patterns to novel verbs in the Spanish present tense inflection (cf. Albright 2008a), are to be considered interparadigmatic analogies. Likewise, the Sanskrit problem at hand is a case of interparadigmatic analogy: some lexical items show a mapping STRONG STEM $C_1aC_1\bar{a}C_2$- : WEAK STEM $C_1eC_2$- that replaces an earlier mapping STRONG STEM $C_1aC_1\bar{a}C_2$- : WEAK STEM $C_1aC_1C_2$-.

The first question to answer before attempting to model an interparadigmatic analogy is: what is the paradigmatic BASE form that serves as the point of departure for rules mapping to other paradigmatic forms? For the Sanskrit perfect, I take the 3.sg.act.ind. as the optimal base, following the criteria generally established by Albright: a good base shows few phonological neutralizations, may show characteristics that cannot be regularly predicted by rule, and often has relatively high token frequency. I lack a precise count, but I assume that the 3.sg.act.ind. is indeed the most frequently attested form of the perfect, and the following phonological characteristics make it preferable:

- In the 2.sg., the inflectional ending *-tha* conditions phonological neutralizations.
- In the weak stem, in biconsonantal roots at least, phonological neutralizations may occur due to regressive assimilation when the two consonants are in contact, e.g., *saśc-* to $\sqrt{sac}$.

Therefore, I will take the form of the stem as does or would appear in the 3.sg.act.ind. as the base for the prediction of the weak stem.

**Sorting the Data**

The objective in this learning simulation is to establish the CONFIDENCE for different pattern mappings between the perfect strong stem and perfect in Vedic biconsonantal roots, for a stage of Vedic prior to the extension of $C_1eC_2$- weak stems beyond the stems *sed-*, *yet-*, and *yem-*. This objective first requires as complete an assemblage as possible of the set of perfect stems that existed in the earliest Vedic (i.e., shortly before the composition of the RV Family Books). To that end, I first gathered a list of all roots of the form $C_1\check{a}C_2$ (in the Sanskritist sense) that attest a perfect stem in Sanskrit from Kümmel 2000 and Whitney 1885 [1963]. From that list, I excluded all perfects attested only in Classical Sanskrit or only reported by grammarians. I then further excluded all perfects for which no verbal form is reported prior to Epic Sanskrit. The group under consideration thus consists of perfects attested at any period of Vedic Sanskrit, and those perfects first attested in Epic Sanskrit, but for which a verbal

| Strong Stem | Weak Stem | Example | Number of Types |
|---|---|---|---|
| $C_1aC_1\bar{a}C_2$- | $C_1aC_1C_2$ | *cakār-* → *cakr-* | 55 |
| $C_1aC_1\bar{a}C_2$- | $C_1eC_2$ | *sasād-* → *sed-* | 3 |
| uv$\bar{a}C_2$- | $\bar{u}C_2$- | *uvāc-* → *ūc-* | 4 |
| $C_1aC_1\breve{a}C_2$- | $C_1i/uC_1i/uC_2$- | *tatar-* → *titir-* | 3 |
| $C_1\breve{a}C_1\breve{a}C_2$ | $C_1\breve{a}C_1\breve{a}C_2$ | *rarābh-* → *rarabh-* | 8 |

Table 8.5: Type Frequency Patterns of Strong Stem : Weak Stem in Vedic Perfects to $C_1\breve{a}C_2$- Roots in Bartholomae-Based Simulation

root can be established at an earlier period.[16] I further excluded all perfects that appeared to inflect only in the middle voice, since the perfect would then consist solely of a weak stem that could not be synchronically derived by mapping from the strong stem (i.e., there would be gaps in the paradigm). I also excluded the perfects to the roots $\sqrt{naś}$ 'attain' and $\sqrt{bhṛ}$ 'bear', which show formal peculiarities that cannot be generalized to the form of any other perfect. Finally, I set aside a small number of other forms for which an old perfect seemed unlikely to me: for instance, although $\sqrt{rāj}$ 'be kingly; shine' attests a number of verbal forms in the RV, and appears with reasonable frequency, the very absence of a perfect until Epic suggests to me that the creation of a perfect there is an innovation.

The assembled learning set thus consisted of 73 input-output pairs, reflecting all $C_1\breve{a}C_2$ that plausibly possessed both a strong stem and a weak stem. In case either the strong stem or the weak stem is not directly attested, I have reconstructed it. In the case of roots such as $\sqrt{sah}$ 'overcome' and $\sqrt{rabh}$ 'seize' that show a $C_1eC_2$- form alongside a weak stem **not** of the from $C_1aC_1C_2$- (e.g., 3.sg.perf.mid.ind. *rārabhe*) in the RV, I used the latter option; this keeps the $C_1eC_2$- outputs precisely *sed-*, *yet-*, and *yem-*. For many perfects for which only a weak stem of the form $C_1eC_2$- is attested, I reconstructed a weak stem of the form $C_1aC_1C_2$-. The major patterns found in this set of input-output pairs are illustrated in Table 8.5. The final row of the table reflects a diverse array of isolated patterns that are limited to one or two types, and in some cases, single tokens; these peculiar forms are entirely substituted in their weak stem by $C_1eC_2$- by Epic Sanskrit, insofar as a perfect to that root is still attested at all in that period of Sanskrit.

**Encoding Forms**

One further and crucial detail requires comment: in order for the $C_1eC_2$- pattern to be learnable at all, one must assume that the structural change of the morphological rule allows for abstract generalization over phonological features. That is to say, the structural de-

[16] This latter decision, namely, to infer the existence of some perfect stems for early Vedic that are not directly attested, is obviously an operational heuristic that is needed to try to reconstruct the actual population of Vedic perfects as best as possible, despite the inevitable gaps in attestation. This procedure likely does correctly infer the existence of some perfect stems that, by accident, have a late direct attestation, though it might also posit some anachronistic forms. Perhaps a classification model based on constellations of verbal forms and lexical semantic features might serve as a more accurate means of determining whether an unattested perfect stem is likely to have existed in early Vedic.

| Variable | Example Encoding | Actual Form |
|---|---|---|
| W = i/u$C_1$ | tWir | *titir-* |
| X = a$C_1$ | sXAd | *sasād-* |
| Q = ā$C_{[+back]}$ | cQan | *cākan-* |
| Y = a$C_{[+back]}$ | cYam | *cakam-* |
| Z = ā$C_1$ | dZAr | *dādar-* |

Table 8.6: Encoding Used for MGL Input Files of Perfect Weak Stem Simulations

scription admits of rules such as $C_{[+nasal, -cor]} \rightarrow C_{[-nasal, -cor]}$. Theoretically, this assumption is unproblematic; *SPE*-style re-write rules regularly admit of generalization over the inputs to a rule. For reasons that are not known to me, the MGL generalizes over sets of features solely in the structural description of the rules that it learns.[17] For example, given the two input-out mappings *tatād-* : *ted-* and *sasād-* : *sed-*, an analyst could easily posit a structural change $aC_{[+cor, -voi]}\bar{a} \rightarrow e$. The MGL however, will not learn such a rule; it will note only two separate possible rules *atā* → *e* and *asā* → *e*. In short, the MGL learns only strict segmental changes as structural changes. Consequently, the linguist aware of potential generalizations over the inputs must encode those generalizations in the input files to the MGL through the use of a variable. For instance, I used the variable X to represent a sequence aCā in the input to the problem at hand; thus, the MGL has the potential to learn a rule X → e. The full set of variables used to circumvent this problem is listed in Table 8.6. Furthermore, because the MGL cannot learn discontinuous morphological changes, the inflectional ending *-a* of the 3.sg. perfect is omitted from inputs, though the inputs are understood to otherwise reflect 3.sg. forms.

At the practical level, the concern is to determine the extent to which a rule XA (= $aC\bar{a}$) → e (generating $C_1eC_2$- forms) can supersede a rule A → ∅ (generating $C_1aC_1C_2$- forms).

#### 8.3.3.2 Analysis of MGL Results

From an input set containing the mappings *sasāda* → *sed-*, *yayāta** → *yet-*, and *yayāma* → *yem-*, the following three minimal rules are learned:

1. aCā → e / s_d;
2. aCā → e / y_$C_{[-approx, -cont, -s.g., -strid, -lat, -dor, -high, -low, -front, -back]}$;
3. then generalizing over those two rules: aCā → e / $C_{[+cont, -nas, -sg, -lab, -lat, -lo, -back, -long]}$_$C_{[-approx, -cont, -sg, -strid, -lat, -dor, -high, -low, -front, -back]}$

The first rule on its own is based on a single form, and thus in principle cannot extend to cover new forms; moreover, the MGL does not consider the reality of such a rule. The

[17] Bruce Hayes (p. c.) informs me that this limitation of the MGL is not intentional; the development of the learner simply did not reach that stage.

second rule is an ISLAND OF RELIABILITY (reliability $1$, confidence $.57$ given the scope of $2$), but which nevertheless has lower confidence than many various versions of the rule $\bar{a} \rightarrow$ Ø, which generates $C_1aC_1C_2$- forms. The more general third rule would also take scope over the strong stems *śaśā́pa**, *sasā́pa**, and *rarā́ma**, and *sasāna**, but it is a fairly weak rule (reliability $0.428571429$ $(3/7)$, confidence $0.3025$ given the scope $7$).

Morphologically, this simulation predicts the existence of, at most, seven possible perfect weak stems with *e* vocalism: *yet-*, *yem-*, *rem-*, *śep-*, *sed-*, *sen-*, and *sep-* – the crucial binding feature is that the $C_1$ of these roots is [+cont]. No morphological rule able to generate a $C_1eC_2$- form even takes scope over inputs such as *babhāja* or *papāca*, to produce desired *bhej-* and *pec-*. The problem lies in the fact that the most general pattern that can be extracted from *sed-*, *yet-*, and *yem-* requires that $C_1$ be a continuant. Moreover, versions of the rule $\bar{a} \rightarrow$ Ø, under this simulation, have consistently higher confidence than $aC\bar{a} \rightarrow e$, and consequently the simulation predicts, morphologically, *sasd-*, *yayt-*, and *yaym-* (all of which would nonetheless surely surface with *e*) as more likely than *sed-*, *yet-*, and *yem-*.

The results of this simulation demonstrate decisively that, were *sed-*, *yet-*, and *yem-* indeed the only perfect weak stems of the $C_1eC_2$- type at some point in the history of Vedic, then the analogical extension of the type would have been well-nigh impossible.[18] Not only is the number of types presumed to form the core of the analogy too small, but those few types form a phonologically closed class. The clear remedy to this situation is to construct a broader basis for the analogy: more non-analogical stems exhibiting a surface form $C_1eC_2$- must have existed than is traditionally assumed.

## 8.4 On the Trail of Vedic Phonotactics

While both Lubotsky's and Bartholomae's assumed bases for the analogical extension of $C_1eC_2$- weak stems clearly starve from a paucity of instantiating types, Lubotsky and Kümmel both note that the expected reduplicated weak stems ($C_1aC_1C_2$-), in many cases, might contain a strange or rare sequence of consonants for Sanskrit. Kümmel (2000: 19) suggests that "diese *e*-Regel wurde ausgenutzt, um seltene Konsonantengruppen zu vermeiden",[19] while Lubotsky (2013: 178) explicitly states that "the clusters $^{\star}$*bhj* [as expected for $^{\text{X}}$*babhj-*], $^{\star}$*pc* [as expected for $^{\text{X}}$*papc-*], $^{\star}$*śk* [as expected for $^{\text{X}}$*śaśk-*], etc. are phonotactically inadmissible in Sanskrit."

What Lubotsky and Kümmel have in mind seems to be a sort of grammatical organization in which the violation of a phonological constraint triggers repair by the application of an alternative morphological process. A combination of morphological and phonological

[18] In this particular case, I assume that the relevant factors behind the analogy are entirely grounded in phonological structure, hence making the MGL a convenient tool for study of the problem. Were one to include morphosyntactic or semantic factors in addition to phonological factors, then one could employ techniques (such as ground the TiMBL model, for instance) to sort the likely class membership of the forms. Since all the forms under consideration here are uniform in their morphosyntactic profile, and lexical semantics seems indiscriminate in this case, the decision to operate soley with phonological factors seems justified.

[19] "This *e*-Rule was employed in order to avoid rare consonant clusters."

constraints with the ranking *Bad Phonotactics ≫ Use-Reduplication ≫ Use-$C_1$e:$C_2$ could predict this behavior.[20] Where reduplication produces no phonotactic violations, Use-$C_1$e:$C_2$ being the lowest-ranking constraint, reduplication surfaces; where reduplication would result in a violation of higher ranking *Bad Phonotactics, a violation of Use-Reduplication instead reflects the optimal candidate (55.a):

(56)

| sad, Perfect, 3.pl. | | *Bad Phonotactics | Use-Reduplication | Use-$C_1$e:$C_2$ |
|---|---|---|---|---|
| a. | ☞ se:dur | | ∗! | |
| b. | sasdur | ∗! | | ∗ |

More fine-grained phonotactic constraints could predict the simultaneous existence of *sed-* alongside *babhj-*, if the rankings *[sd] ≫ Use-Reduplication but Use-Reduplication ≫ *[bhj] were to hold. The "analogical" extension of the $C_1eC_2$- forms would then be not a morphological change *per se*, but a progressive change in the relative rankings of specific phonotactic markedness constraints and the morphological constraint Use-Reduplication.

Problematic for such an analysis is that $C_1eC_2$- forms also appear to roots where the sequence of $C_1C_2$ otherwise appears to be licit in Vedic, e.g., *sep-* (for $^{X}$[sasp-] — [sp] is a licit sequence in Vedic in both heterosyllabic and onset parses). Therefore, it is necessary to assume some purely morphological application of a $C_1eC_2$-Rule, in case the competing reduplicated candidate cannot be excluded on phonotactic grounds. Besides questions as to whether such an interaction between morphological constraints and phonotactic constraints is even a desirable theoretical prediction, such a model is plainly inadequate for the case at hand.

Nevertheless, the observation that at least some further roots belonging to the class of $C_1eC_2$- forms would have contained a phonotactically ill-formed sequence in a reduplicated stem $C_1aC_1C_2$- provides an important clue that can drive this investigation forward. If morphologically generated underlying /bhəbhj-/ could not surface faithfully as such, is then *bhej-* the correct repair? The answer seems to be "yes": Ved. *bhej-* could result from PIIr. *[$b^h$əːj́-],[21] by deletion and compensatory lengthening, to repair the phonotactically illicit /$b^h$j́/. The problem then becomes merely how to confirm or discover exactly which sequences of segments would have produced fatal violations of phonotactic constraints in Vedic.

### 8.4.1 Phonotactic Learning

As a first-pass, informal attempt at the evaluation of a language's phonotactic grammar, we may simply make anecdotal observations concerning the co-occurrence frequency of segments in some corpus. Intuitively, segments that occur together often would constitute

[20] See now studies on the formalization of allomorphy phenomena in Bonet et al. Forthcoming and similarly Mascaró et al. 2007.

[21] In Sandell 2014a, I have argued that PIIr. *[əz] and *[əž] regularly give PInd. [əː] and then ultimately Vedic *e*.

entirely well-formed sequences, while sequences that never or very rarely occur might be avoided or altogether banned. In applying this method to electronic versions of the RV and Franceschini 2007, I find no instances of *bhj* (as in a $^{X}$*babhj-*) or *śk* (as in a $^{X}$*śaśk-*); *pc* (as in a $^{X}$*papc-*) does not occur, although *pch* does rarely, in the compounds *triṣṭupchandas-* '*triṣṭubh-* meter' and *anuṣṭupchandas-* '*anuṣṭubh-* meter'. Similarly, *sh* (as in a $^{X}$*sash-*) is totally absent, and *śp* is also very rare, appearing only in the compound *viśpati-* 'protector of the village; lord' and the personal name *viśpāla-*. In general, combinations of labials or velars with palatals appear to be phonotactically problematic for Sanskrit.

More ideal, however, would be a means not only to observe seeming phonotactic absences, but to predict, based on the actually occurring combinations of sequences in the language, what potential word forms could pose phonotactic problems, and how severe those problems might be. In short, we want to be able to predict the well-formedness of some sequences that happen not to occur by accident, and have a precise motivation to exclude other ill-formed combinations. Hayes and Wilson (2008) have written software, the UCLA PHONOTACTIC LEARNER, that fulfills precisely this desideratum.[22]

Hayes & Wilson's learner employs a maximum entropy (MaxEnt) model (see generally Manning and Schütze 1999: Ch. 16 or Goldwater and Johnson 2003) to assign weights to markedness constraints. The maximum entropy model has the effect of maximizing the probability of observed forms and minimizing the probability of just those unobserved forms that "differ in a principled way from the observed forms" (Hayes and Wilson 2008: 385). This specific learning method begins by constructing every possible constraint (from two to four segments in length) from the natural classes within the features of a segment inventory. However, because this step produces an unwieldy number of constraints for which to calculate the appropriate weights, Hayes & Wilson employ a heuristic to limit the constraint set: only constraints meeting a given accuracy threshold are maintained in the grammar. Accuracy here is defined as the number of violations of a constraint observed in the raw data divided by the expected number of violations in a given grammar. Constraints are gradually added to the grammar using a weaker accuracy criterion until reaching either a user-provided accuracy limit or a user-provided maximum number of constraints. At each step, the learner initially assigns a weight of 1 to all constraints. From a representative sample of the language, the learner employs the Conjugate Gradient algorithm (Press et al. 1992) to iteratively converge on the constraint weights that best maximize the probability of observed forms (i.e., the global maximum of the search space).

I treated here the RV *saṃhitāpāṭha*[23] as training data for Vedic phonotactics; I prepared the appropriately formatted input file using a Python script, and made an appropriately formatted file of phonological features based on the file of features prepared for the MGL simulations.[24] I allowed the UCLA Phonotactic Learner to acquire and train a maximum of 130 constraints. I then fed the resulting grammar a list of 3.pl.perf. forms, both actually occurring and nonce, of the form $C_1aC_1C_2$-, for every perfect that attests a $C_1eC_2$- form and was used in

[22] Software available from:
`http://linguistics.ucla.edu/people/hayes/Phonotactics/Index.htm`.

[23] I.e., the text of the RV that shows all surface sandhi.

[24] Available here: `https://github.com/rpsandell/SandellDiss`.

| Form | Score | MaxEnt Value | Constraints Violated |
|---|---|---|---|
| *cacmur* | 9.946 | $4.8 * 10^{-5}$ | ⋆[-continuant,-voice,-anterior][+consonantal,-approximant,+voice]; |
| | | | ⋆[-continuant,-voice,-spread glottis,+front][-dorsal] |
| *cacrur* | 4.901 | $7.4 * 10^{-3}$ | ⋆[-continuant,-voice,-spread glottis,+front][-dorsal] |
| *jajpur* | 11.487 | $1 * 10^{-5}$ | ⋆[-sonorant,+voice][-voice]; |
| | | | ⋆[+voice,+dorsal][-sonorant,-spread glottis,-dorsal]; |
| | | | ⋆[-approximant,+voice,+coronal][-voice,-coronal] |
| *dad.hur* | 10.017 | $4.5 * 10^{-5}$ | ⋆[-approximant,+voice,-lateral][-sonorant,+continuant]; |
| | | | ⋆[+voice,+anterior][-sonorant,+continuant] |
| *papcur* | 6.794 | $1.1 * 10^{-3}$ | ⋆[-continuant,-dorsal][-approximant,-anterior]; |
| | | | ⋆[-continuant,-dorsal][-approximant,+dorsal] |
| *pabdur* | 3.318 | $3.6 * 10^{-2}$ | ⋆[+voice,-coronal][-nasal,-spread glottis,+anterior] |
| *babhjur* | 6.794 | $1.1 * 10^{-3}$ | ⋆[-continuant,-dorsal][-approximant,-anterior]; |
| | | | ⋆[-continuant,-dorsal][-approximant,+dorsal] |
| *śaśkur* | 8.249 | $2.6 * 10^{-4}$ | ⋆[-nasal,+dorsal][-approximant,-labial,-coronal]; |
| | | | ⋆[+consonantal,+front][-approximant,-labial,-coronal] |
| *śaśpur* | 5.154 | $5.8 * 10^{-3}$ | ⋆[-coronal,+dorsal][-sonorant,-continuant,-anterior] |
| *sasdur* | 9.234 | $9.7 * 10^{-5}$ | ⋆[-voice][-sonorant,+voice]; |
| | | | ⋆[-voice,-labial,-dorsal][-nasal,+voice,+anterior] |
| *sashur* | 7.398 | $6.1 * 10^{-4}$ | ⋆[-voice][-sonorant,+voice] |

Table 8.7: Potential Perfect Weak Stems and Phonotactic Constraint Violations

the MGL simulation in 6. Figure 8.7 gives the penalty score (harmony), MaxEnt value (effectively the form's probability = $e^{-Score}$), and the constraints violated for several nonce forms. Note also that in all of these forms the violations incurred stem solely from the sequence of two consonants.

Inspection of results in the empirical tests of Hayes and Wilson 2008 suggests that a score above $4$–$4.5$ is fatally bad. This rule of thumb accords with the rarity of [bd] (occurring in forms such as *upabdi-* and *āpibdamānaḥ*), and the total absence of most other sequences, except in compounds in a couple of cases. One should therefore conclude that an "underlyingly reduplicated"[25] PIIr. 3.pl.perf. */pə-pč-r/ or */bʰə-bʰǰ-r/ would be repaired to *[pəːčr̥] and *[bʰəːǰr̥].[26] We can therefore conclude that the perfect weak stems *cem- cer-*, *jep- deh-*, *bhej-*, *śek-*, *śep-*, and *seh-* likely arose as repairs to phonotactic violations, and thus are not analogical at all.[27] For the moment, we may continue to assume that *sed-*, *yet-*, and *yem-* result from regular sound changes to PIIr. *[səzd-], *[i̯əi̯t-], and *[i̯əi̯m-], since the phonotactic grammar constructed here identifies no problems with the sequences [i̯t] or [i̯m], and I presume that [zd] well-formed in PIIr. (cf. Ved. impv. *dehi* 'give!' < *[dəzdʰí]; the badness

[25] I here abstract away wholly from the question of how the reduplicant is generated. Note that a form like [pəːc-], would show gross misalignment of the left edge of the base with the left edge of the reduplicant, if analyzed within Base Reduplicant Correspondence Theory (McCarthy and Prince 1995).

[26] No Iranian evidence directly confirms or denies the Proto-Indo-Iranian status of such forms.

[27] Whether perfects to the roots $\sqrt{cam}^{i}$ 'sip' and $\sqrt{jap}$ 'whisper' existed in early Vedic is less certain, but are included following the selection procedure described at **8.3.3.1** above.

of [sd] above resulting from the fact that obstruent clusters that disagree in voicing are illicit in Vedic).

### 8.4.2 Further Simulation and Further Failure

Departing from the file prepared with the simulation described in 8.3.3.1, the input-output mappings were changed to reflect the new evidence from phonotactics, e.g., input *babhājа* now maps onto *bhej-* rather than *bhabhj-*, as in the first simulation. A total of 12 types then instantiate some version of a rule $aC\bar{a} \rightarrow e$ in the training data for this simulation, rather than just 3. Given this now larger and phonologically more diverse number of types belonging to the $aC\bar{a} \rightarrow e$ class, it is more likely that a rule could take scope of an input that originally provided support to the $\bar{a} \rightarrow \emptyset$ class; iteratively running "generations" of the MGL, with corresponding adjustments to input-output pairs to reflect new winners, might then capture the gradual expansion of the $C_1eC_2$- type at the expense of the $C_1aC_1C_2$- type.

The first generation learns the following winning minimal rules $aC\bar{a} \rightarrow e$:

1. aCā → e / $C_{[\text{-approx, -son, -nas, -ant, -lat, -low, -back}]}$_$C_{[\text{+cons, -approx, -cont, -s.g., -ant, -strid, -lat, -low, -back}]}$ is an Island of Reliability covering *cem-*, *jep-*, *pec-*, *bhej-*, *śek-*, and *śep-* (reliability 1, confidence .852 with scope of 6).

2. aCā → e / $C_{[\text{-nas, -s.g., -lab, -ant, -lat, +dor, +hi, -low, +front, -back}]}$_$C_{[\text{-approx, -cont, -s.g., -cor, -ant, -strid, -lat, -lo, -front, -back}]}$ is an Island of Reliability covering *cem-*, *jep-*, *yem-*, *śek-*, and *śep-* (reliability 1, confidence .825 with scope of 5).

3. aCā → e / $C_{[\text{-approx, -son, -nas, -s.g., -lab, +cor, -lat, -low, -back}]}$_$C_{[\text{-approx, -son, -nas, -lab, -strid, -lat, -front, -back}]}$ is an Island of Reliability covering *deh-*, *śek-*, *sed-*, and *seh-* (reliability 1, confidence .786 with scope of 4).

4. aCā → e / $C_{[\text{+cont, -nas, -s.g., -lab, -ant, -lat, +dor, +high, -low, +front, -back}]}$_$C_{[\text{-approx, -cont, -s.g., -strid, -lat, -low, -front, -back}]}$ is an Island of Reliability covering *yet-*, *yem-*, *śek-*, and *śep-* (reliability 1, confidence .786 with scope of 4).

5. aCā → e / $C_{[\text{-approx, -son, -cont, -nas, -voice, -s.g., -ant, -strid, -lat, -low, -back}]}$_$C_{[\text{-s.g., -ant, -strid, -lat, -lo, -back, -long}]}$ is an Island of Reliability covering *cem-*, *cer-*, and *pec-* (reliability 1, confidence .0.718, with scope of 3).

While these rules suffice to maintain all of the original input-output mappings that instantiate the *e*-rule, no other rules with broader scope have sufficient confidence to clearly replace any of the $C_1aC_1C_2$- outputs. For instance, this simulation predicts the stem *sasp-* based on a rule with reliability 1 and confidence of .872 with scope of 7, while the best rule generating a competing *sep-* has a reliability of .8 and confidence of .61 with scope of 5. It is thus not evident that the $C_1eC_2$- pattern should expand at all. At face value, the results of this simulation suggest that the type ought to have remained restricted to a small set of forms whose perfect weak stem would have been phonotactically unacceptable.

However, if some forms were to convert to the $C_1eC_2$- type, the effect would be to give certain *e*-rules higher confidence than the confidence of the winning $\bar{a} \rightarrow \emptyset$ rule. For instance, a rule with a scope of 7 would generate **tatp*- and **dadbh*-, but a competing *e*-rule with scope 9, to which *tatp*-, *dadbh*-, and *sasp*- are the three exceptions, would have higher confidence were the forms to instead become *tep*-, *debh*-, and *sep*-. Since rules with broader scope are, by definition, more general, one might be willing to see the conversion of **tatp*- >> *tep*- as indicative of a bias towards more general (i.e., supraminimal) rules. Specifically, the initial analogical extensions of the $C_1eC_2$- type could reflect the tendency for "automatic overgeneralization" in morphological acquisition reported by Kapatsinski (2013: 124–30).

Although varieties of "simplicity bias" are reported in the literature on phonological acquisition (cf. citations in Moreton 2012), one evident disadvantage to assuming such a bias in this case is that it vitiates the power of islands of reliability. If maximally general rules were universally preferred, many more English preterites would be expected to succumb to the general *-ed*-rule, rather than sustaining their ablaut patterns supported by islands of reliability (cf. Albright and Hayes 2003). The question then becomes: under what precise conditions can morphological rules of lower confidence attract forms because of their greater generality? For the particular case at hand, the *e*-rule is still not the best supported rule, given that the Ø-rule claims more types; thus, in terms of maximal generality, the *e*-rule has no advantage whatsoever over the Ø-rule.

Furthermore, it is not clear that creating *e* forms via "overgeneralization" really does serve to simplify the grammar. The Ø-rule for weak stems is patently the more widespread and reliable means of forming perfect weak stems – it applies not only to more roots of the shape $C_1aC_2$, but also to any root containing a sonorant between two obstruents that can serve as a syllable nucleus (e.g., *cikit*- to a root /cəjt-/). Hence, the most sensible interpretation of "overgeneralization" is for the application of the Ø-rule wherever phonotactically permitted. In turn, *e*-rule extensions must be motived by the power of Islands of Reliability. I therefore conclude that the results from this simulation, employing an expanded set of $C_1eC_2$- forms, still directly fails to predict new forms. Moreover, I see no other likely interpretation under which these results could be interpreted as leading to analogical extensions.

## 8.5 OCP-Syllable in Indo-European Reduplication

The failure of the preceding simulation to fully account for the Vedic data suggests that we may be placing too much explanatory burden on an analogical process to create the attested forms. In fact, the only forms that truly must (philologically speaking) be explained analogically are those six (*pet*, *ten*-, *med*-, *men*- *sec*- and *ter*-) that actually do directly attest the reduplicated weak stem. The strong converse of this premise (i.e., only those six forms need be explained analogically) is that all the $C_1eC_2$- form in all other cases can be taken for granted; if so, then the analogical creation of those six forms, we will see, is achieved trivially, because base for the analogy is then so strong. Nevertheless, we assume that at least some few other forms, such as *sep*-, are due to an analogy because excluding the historical existence of a PIIr. *[səsp-] is doubtful.

In the preceding section, the examination of Vedic phonotactics already confirmed that a substantial number of roots to which we find $C_1eC_2$- forms would contain a phonotactically dispreferred sequence in a reduplicated $C_1aC_1C_2$- form. While the linear sequence [jp], for instance, may be ill-formed in Vedic, these phonotactic constraints might be epiphenomenal. One might reasonably wonder whether some larger generalization is being missed. In work by Zukoff (2014) on reduplication in Greek, another phonotactic generalization appears in the form of an Obligatory Contour Principle constraint over the domain of the syllable:

(57) OCP-Syllable (OCP-$\sigma$) – Assign one violation mark * for every syllable that contains two identical segments.

Reduplicated formations involving copy of the leftmost root consonant, as are typical of Greek and Sanskrit, can easily lead to OCP-$\sigma$ violations when the leftmost consonant of the base cannot be parsed into the onset of the following syllable. For instance, in Proto-Indo-Iranian in terms, the root $\sqrt{pač}$ 'cook' could be said to morphologically build the 3.pl. of a perfect weak stem as /pə-pč-r/. Given a high ranking of OCP-$\sigma$, however, an output containing a syllable [pəp] would be excluded, while an output [pə.pčr̥], with word-medial onset [pč-], would not. A third possibility, *[pəː.čr̥], with deletion of a /p/ and compensatory lengthening, is to be expected just in case both OCP-$\sigma$ and specific phonotactic constraints that would militate against an onset [pč-] outranking Max-C. Since a surface form *[pəː.čr̥] indeed can be directly continued in Vedic as *pecur*, and comparison to other Indo-European languages gives little reason to believe that an onset *[pk$^{w}$] would have been licit in PIE, we should reconstruct *[pəː.čr̥] as the Proto-Indo-Iranian output of /pə-pč-r/, as represented in the following tableau.[28]

(58) OCP-$\sigma$ and *Bad Phonotactics block faithful outputs

[28] Worth considering on this point is Vedic *kṣumánt-* 'rich in food', Av. *fšumant-* 'having (wealth in) cattle', which continues PIIr. *[pšumánt-] < PIE Transponat *[pḱu-mént-], and other similar forms (e.g., Ved. *purukṣú-* 'rich in food') where the PIIr. labial+palatal sequence (given a UR */pću-mant-/) has appears to have been repaired to [pš-] without deletion. Perhaps the difference lies in the fact that outright deletion of /p/ in this case would have left that /p/ without any surface exponent (even its weight or timing). The ranking of possible repairs to */pć/ is then: delete /p/ (with some feature preservation, e.g., compensatory lengthening) ≫ dissimilate /ć/ to [š] ≫ delete /p/ (without any feature preservation). Alternatively, a sound change IE *[pḱV] > PIIr. *[pšV] is at work (cf. Cathcart 2012: 31–2), and we assume that *[pšV] was not phonotactically illicit in PIIr. Ved. *virapśá-* 'abundance' < *[u̯īrapću̯á-] ← */u̯ih$_1$-ro-pḱ-u̯-ó-/ is troubling, however: the sequence *[pć] remains undisturbed. This example implies that a change effecting *[pḱV] or repair to /pć/ targets specifically that sequence when it would unavoidably be parsed into an onset.

| /pa-pč-r/ | | *Bad Phonotactics | OCP-σ | Max-μ | Dep-V-IO | Max-C-IO |
|---|---|---|---|---|---|---|
| a. | ☞ pəː.čr̥ | | | | | * |
| b. | pəp.čr̥ | | *! | | | |
| c. | pə.pčr̥ | *! | | | | |
| d. | pə.pə.čr̥ | | | | *! | |
| e. | pə.čr̥ | | | *! | | |

Candidates b. and c. respectively violate the high-ranking markedness constraints OCP-$\sigma$ and *Bad Phonotactics; candidate d. opts for a repair via vowel epenthesis, but which is excluded by the ranking Dep-V-IO $\gg$ Max-C-IO. Candidate e. is meant to reflect the fact that compensatory lengthening is expected in the output, though I will not worry here about the specific phonological problems involved in generating this compensatory lengthening.[29]

In effect, under conditions where phonotactic or other constraints on licit syllable structure rule out certain onset parses, OCP-$\sigma$ then further excludes the option of a segment's faithful realization in a coda, given that the resulting syllable would contain two identical segments. Particularly notable is that, in any root of the shape $C_{[+son]}aC_{[-son]}$, (e.g., $\sqrt{rabh}$ 'seize'), a faithful form with clear reduplication could not surface: a *[rə.rbh-] would contain an onset with a gross sonority sequencing violation ($C_{[+son]}C_{[-son]}$ onsets categorically do not exist in old Indo-European languages) and a *[rər.bh-] would violate OCP-$\sigma$. For the present problem, the introduction of OCP-$\sigma$ has the potential to obtain yet more $C_1eC_2$-forms phonologically, and perhaps to enlarge the set sufficiently such that new analogical forms might take hold.

### 8.5.1 Motivating and Situating the Role of OCP-$\sigma$

Before proceeding in further analysis of the problem at hand, two brief digressions are needed. The first concerns the plausibility of a constraint like OCP-$\sigma$. Typologically, consonant concurrence restrictions, are abundantly attested: see, for instance, Frisch et al. 2004, Pozdniakov and Segerer 2007, or Coetzee 2014, and other literature cited in those articles. Whether constraints like OCP-$\sigma$ must be distinguished from constraints on immediate adjacency, like the standard OCP, is not wholly certain; Coetzee suggests that consonantal concurrence restrictions can be framed in autosegmental terms as constraints on adjacency of certain features on the consonantal tier. At present, I will assume that there is probably no need to separate OCP-$\sigma$ from the broader effects of the Obligatory Contour Principle in old Indo-European languages and PIE.

Sanskrit, however, allows for morphologically derived geminate consonants, unlike PIIr. or PIE; cf. forms such as 2.pl.pres.impv. *dattá* ← /də-daː-tá/ (to $\sqrt{dā}$ 'give'). If we generally equate OCP-$\sigma$ and the OCP generally, then we should not expect to find synchron-

[29] The constraint PositionCorrespondence proposed by Topintzi (2010: 123), which forces each underlying segment to correspond on the surface to either a root node or a mora, would, I think, be adequate for this case. For discussion of the same problem with respect to Germanic forms, see Sandell and Zukoff 2014.

ically operating OCP-$\sigma$ effects in Vedic. The metrical treatment of *muta cum liquida* (i.e. [VC$_{[-son]}$C$_{[+son]}$V]) sequences in Sanskrit indicates that they always parse heterosyllabically. For example, in a reduplicated aorist form such as 3.sg. *adudrot* (to $\sqrt{drav}$ 'run', RV 2.30.3c, *triṣṭubh* meter, in the pāda's cadence), the scansion is ⏑ — —, thus indicating a syllabic parse [ə.dud.roːt]. Although C$_{[-son]}$C$_{[+son]}$ onsets are abundant word-initally, and thus clearly licit in general in Sanskrit, their avoidance word-internally, even where syllables containing identical segments would result, establishes that a constraint like OCP-$\sigma$ has no truly active *synchronic* effect in Vedic. The irrelevance of the OCP for the synchronic phonology of Vedic directly impacts the current issue in two ways:

1. First, forms like *pec-* or *rebh-* are rendered phonologically opaque. A learner would precisely need a sufficiently high-ranked OCP in order to recover [peːc-] and [reːbh-] as reflecting /pə-pc-/ and /rə-rbh-/. Consequently, the learner's account for *pec-* and *rebh-* comes to be situated in the morphological component of the grammar. Such morphologization creates precisely the conditions for analogical extension proper.

2. Forms that have a metrical scansion that implies an OCP-$\sigma$ violation in Vedic, like *paptur* ([pəp.tur]), metrically scanning — —, could easily be prosodified in a grammar that permits OCP violations, given that the learner inherited a string [pəptur] as a learning datum. The synchronic existence of [pəp.tur], with that syllabification, does not necessarily indicate that the same syllabification existed at an earlier historical period (i.e., in PIIr. or PIE). To be explicit: an earlier PIIr. grammar could have generated [pə.ptr̥] from /pə-pt-r/, but a Vedic learner, then presented with the datum [pəptur], and no reason to believe that a syllable [pəp] is ill-formed (because his OCP constraint is low-ranked), assigned the parse [pəp.tur] to that string in preference to [pə.ptur] (which would contain a complex onset without rising sonority).

Further evidence is available to suggest that the an even stronger version of OCP-$\sigma$ than that proposed by Zukoff 2014 may have been present in PIE and PIIr. This evidence comes principally from the underrepresentation of certain PIE root shapes. Roots of the shape */C$_1$eC$_2$-/, in which the two consonants share both place and manner of articulation, are virtually non-existent: with dentals, there are no roots **tet-*, **ted-*, **ted$^h$-*, **det-*, **ded-*, **ded$^h$-*, **d$^h$et-*, **d$^h$ed-*, or **d$^h$ed$^h$-* A difference in glottal state alone (voicing, +/- spread glottis) may be insufficient to avoid the OCP violation. The stronger version of OCP-$\sigma$ could then be formulated as in (54):

(59) OCP-Syllable (OCP-$\sigma$) – Assign one violation mark * for every syllable that contains two segments identical in place and manner of articulation and adjacent on the consonantal tier.

Given a phoneme inventory of $25$ consonants in PIE, there are 625 possible permutations of any two elements, allowing for repetition ($= 25^2$). $55$ possible permutations, allowing for repetition, would violate OCP-Syllable. Based on the reverse index of Rix et al. 2001, there

is just one root that violates either Zukoff's formulation in (52) or my formulation in (54) (**ses-*),[30] but 180 distinct biconsonantal roots:

| | OCP-violating | All |
|---|---|---|
| Attested | 1 | 180 |
| Possible | 55 | 625 |

A $\chi^2$ contingency test indicates that the difference between the OCP-violating set and the total set is highly significant: $\chi^2 = 12.1389$, $p < .001$. In roots of the shape /CReC/ or /CeRC/, I count a total of 9 OCP-SYLLABLE violating roots:[31]

| | OCP-violating | All |
|---|---|---|
| Attested | 9 | 465 |
| Possible | 660 | 7500 |

I therefore believe that there is compelling statistical evidence to reconstruct a constraint for PIE as stated in (54).

### 8.5.2 A Feasible Model and Chronology of $C_1eC_2$- Forms

Based on our general knowledge of syllabification in PIE (on which see Byrd 2015), the best question to ask now is: how many $C_1aC_2$- roots might have been caught between the rock of OCP-$\sigma$ and hard place of an illicit onset sequence or other phonotactic problem (as diagnosed in 8.4 above)? Table 8.8 lists all roots to which $C_1eC_2$- forms are attested in Vedic or Epic Sanskrit, and considers whether the sequence of that root's two consonants could form a permissible onset in Sanskrit or some pre-stage thereof.

The results evident in that table are striking. Note especially that roots of the form $\sqrt{C_{[+son]}aC_{[-son]}}$ cannot produce a licit onset [RT-], which would result in a severe SONORITY SEQUENCING PRINCIPLE (Clements 1990) violation. Thus any root of a shape such as $\sqrt{rabh}$ would, in PIIr., inevitably have generated a perfect weak stem *[$C_1$əː$C_2$-]. There is thus a real possibility that the vast majority of Vedic $C_1eC_2$- forms are completely phonological in origin.

Just as importantly, precisely those roots to which we find an attested weak stem $C_1aC_1C_2$- (*tatn-*, *papt-* *mand-*, *mamn-*, and *saśc-*) either would allow an onset parse (PIIr. *[pə.pt-], *[mə.mn-], *[sə.ćč-]), or would not violate OCP-$\sigma$ with a coda parse ([mən.d-], [səć.č-]). Only for an onset [tn-] (as in a presumed parse *[tə.tn-]) is direct evidence lacking, but given the existence of onsets [tm-] (Ved. inst.sg. *tmanā* 'soul, self'), [dm-] (YAv. gen.sg. *nəmō* <

[30] I exclude the root **tetḱ-*, since I take the forms assigned to it in Rix et al. 2001 to be reduplicated forms belonging to the root **teḱ-*.

[31] This figure derives from counting some items as .5, since whether they would be OCP-SYLLABLE violating depends on a part of the reconstruction that is ambiguous. These roots, as given in Rix et al. 2001, are: **tend-* 'schneiden, spalten', **terd-* 'durchboren, spalten', **derd$^h$-* '(ein)schlafen', **d$^h$eu̯d$^h$-* 'erschüttern', **h$_1$reh$_1$-* 'fragen', **h$_3$neh$_2$-* 'genießen', **ĝ$^h$ei̯ĝ$^h$-*/*g$^h$ei̯g$^h$-* 'lechzen', **h$_1$erH-* 'waschen', **h$_2$erH-* 'sich auflösen, verschwinden', **h$_1$eu̯H-* 'helfen, fördern', **h$_2$eu̯H-* '(Fußbekleidung) anziehen', and **k$^{(w)}$eu̯k-* 'sich biegen'.

| Root | Gloss | $C_1C_2$ Onset Attested? | Non-$C_1eC_2$ Weak Stem Attested? |
|---|---|---|---|
| $\sqrt{cam^{(i)}}$ | 'sip' | No | No |
| $\sqrt{car}$ | 'move' | No | No |
| $\sqrt{jap}$ | 'whisper' | No | No |
| $\sqrt{tan}$ | 'stretch' | No[a] | *tatn-* |
| $\sqrt{tam}$ | 'tire' | Yes | No |
| $\sqrt{tap}$ | 'be warm; heat' | No | No |
| $\sqrt{tar^{i}}$ | 'cross over' | Yes ([tr]) | *titir-* |
| $\sqrt{dabh}$ | 'deceive' | No | No |
| $\sqrt{dah}$ | 'burn; extinguish' | No[b] | No |
| $\sqrt{nad}$ | 'sound' | No | No |
| $\sqrt{nam}$ | 'bow' | No | No |
| $\sqrt{naś}$ | 'perish' | No | No |
| $\sqrt{nah}$ | 'bind' | No | No |
| $\sqrt{pac}$ | 'cook' | No | No |
| $\sqrt{pad}$ | 'fall, go' | No [bd-][c] | No |
| $\sqrt{pat}$ | 'fly, fall' | Yes (for PIE and PIIr., not in Vedic) | *papt-* |
| $\sqrt{bhaj}$ | 'divide' | No | No |
| $\sqrt{math^{i}}$ | 'rob' | No [nth] | No (NB no OCP-$\sigma$ violation in *manth-**). |
| $\sqrt{mad}$ | 'exhilarate' | No [nd] | *mand-* (NB no OCP-$\sigma$ violation in *mand-*) |
| $\sqrt{man}$ | 'think' | Yes (for PIE and PIIr., not in Vedic) | *mamn-* |
| $\sqrt{yat}$ | 'take a position' | No | No |
| $\sqrt{yam}$ | 'stretch out; hold' | No | No |
| $\sqrt{rabh/labh}$ | 'grasp' | No | *rărabh-* |
| $\sqrt{raṇ}$ | 'ring' | No | No |
| $\sqrt{ram}$ | 'be content' | No | No |
| $\sqrt{ras}$ | 'roar' | No | No |
| $\sqrt{rāj}$ | 'be kingly, shine' | No | No |
| $\sqrt{lap}$ | 'prattle' | No | No |
| $\sqrt{śak}$ | 'create, shape' | No | No |
| $\sqrt{śap}$ | 'curse' | No | No |
| $\sqrt{sac}$ | 'follow' | Yes ([śc]) | *saśc-* |
| $\sqrt{sad}$ | 'sit' | No | No |
| $\sqrt{sap}$ | 'care for, honor' | Yes [sp] | No |
| $\sqrt{sah}$ | 'overpower, win' | No | *sāh-*[d] |

Table 8.8: $C_1eC_2$- Forms and Well-Formed Onsets

[a] The Younger Avestan forms *ϑnasaṯ* (Pahlavī Vīdēvdād 6.52.0) and *ϑnātō* (Frahang-ī ōīm 656) might suggest the existence of forms with initial *[tn-], but the interpretation of both forms is entirely uncertain; Bartholomae (1904: 799) does not venture to offer glosses.

[b] [dɦ-] is clearly an ill-formed onset in Sanskrit, though a PIE *[de.dg$^{wh}$-] with onset [dg$^{wh}$V-] would have been licit, and would give Ved. *dakṣ-*.

[c] Medial [-bd-] attested in reduplicated *píbdamāna-*.

[d] The stem *sāh-* appears only in the perf.part.act. *sāhvắṁs-* (10× RV), and must itself be a lexicalized archaism, from PIE *[sēĝ$^{h}$u̯os-] ← */se-sĝ$^{h}$-u̯os-/, showing the effects of deletion and CL driven by OCP-$\sigma$.

*[dmés] 'of the house') and [pn-] (Gk. [pnéɔː] 'breathe'), [tn-] would not differ in a principled way.

All of the perfect weak stems to the roots in Table 8.8, in PIIr., insofar as they existed at

all, would therefore have shown a stem *[$C_1$əː$C_2$-], giving Vedic $C_1eC_2$-, with the following exceptions: *[tətn-], *[tətm-], *[tətr̥H-] (> Ved. *titir-*), [pəpt-], *[mənt$^h$-], *[mənd-], *[məmn-] *[səćč-], and *[səsp-]. Precisely the Vedic reflexes of those stems must be explained analogically. When the MGL output forms in my original MGL input file are altered to reflect the further number of forms that likely possessed a $C_1eC_2$- weak stem in early Vedic on account of PIIr. phonology, finally capturing the necessarily analogical forms becomes possible.

In a first generation of learning, the following new forms are predicted:

- *tem-**[32] is predicted on the basis of *cem-*, *jep-*, *tep-*, *debh-*, *deh-*, *nem-*, *neh-*, *pec-*, and *bhej-* (rule with confidence of .79, to which a *tatm-* would be the only exception). The competing rule that would generate *tatm-* has a confidence of .57.
- *sep-* is predicted on the basis of *jep-*, *tep-*, *debh-*, *deh-*, *ned-*, *neh-*, *śek-*, *śep-*, *sed-*, and *seh-* (rule with confidence of .81, to which a *sasp-* would be the only exception). The competing rule that would generate *sasp-* has a confidence of .72.
- *sec-* is predicted on the basis of *jep-*, *tep-*, *debh-*, *deh-*, *ned-*, *neh-*, *pec-*, *bhej- śek-*, *śep-*, and *seh-* (rule with confidence of .73, to which *saśc-* and *sasp-* would be exceptions). The competing rule that would generate *saśc-* is very close in confidence, at .72.

This generation seems to approximately reflect the situation of the RV Family Books, in which we find 3.pl.perf.act.ind. *sepur* (6.29.1a), but *papt-*, *mand-*, and *mamn-*. *saśc-* is found in the Family Books as well, and into Book I, but given the close degree of confidence between the competing rules predicting *sec-* and *saśc-*, this is unsurprising.

In a subsequent generation of learning, altering the outputs to reflect *tem-*, *sep-*, and *śec-*, the following new forms are predicted:

- *ten-* is predicted on the basis of *tep-*, *tem-*, *debh-*, *deh-*, *ned-*, *nem-*, *neh-*, and *lep-* (rule with confidence of .76, to which *tatn-* would be the only exception). The competing rule that would generate *tatn-* has a confidence of .71.
- *ter-* is predicted on the basis of *cem-*, *cer-*, *tep-*, *tem-*, and *pec-* (rule with confidence of .66, to which a *tatr-* would be the only exception). The rule that would generate a competing *tatr-* has a confidence of .65.[33]
- *sen-** is predicted on the basis of *yet-*, *yem-*, *rem-*, *lep-*, *śek-*, *śep-*, *sed-*, and *sep-* (rule with confidence of .79).

This generation predicts the creation of $C_1eC_2$- forms to two original laryngeal-final roots that attest non-canonical weak stems (*titir-*/*tutur-* and *sasan-*) as a result. *ten-* is first attested in the AV, while forms of *tatn-* are found in Book I of the RV, while *pet-*, which is found in later

[32] To $\sqrt{tam}$ 'faint'; the stem *tem-** is accidentally unattested, but trivially predicted under Pāṇini's rule under 8.1.2 above. The same applies to *sen-** below.

[33] As far as I am able to easily discern, the perfect weak stem to $\sqrt{tar^i}$ is unattested between the RV and Epic.

portions of the RV, is not yet predicted. But it should be noted that the confidence of the rule predicting *papt-* at this stage (.825) barely exceeds the confidence of the rule predicting *pet-* (.82). Perhaps both *pet-* and *papt-* were simultaneously viable forms.

In a further generation of learning, altering the outputs to reflect *ten-*, *ter-*, and *sen-*, the following new forms are predicted:

- *pet-* is predicted on the basis of *cem-*, *ten-*, *tep-*, *tem-*, *pec-*, *śek-*, *śep-*, *sec-*, *sed-*, *sen-*, *sep-*, and *seh-* (rule with confidence of .832). The rule generating competing *papt-* remains very close in confidence, at .825.
- *bedh-* is predicted on the basis of *ten-*, *tem-*, *debh-*, *deh-*, *bhej-*, *sed-*, *sen-*, and *seh-* (rule with confidence of .76). The competing rule that would generate *babdh-* has a confidence of .71.

At this stage, the *m*-initial forms remain in Islands of Reliability that predict *mamn-*, *mand-*, and *manth-*. With the availability of *pet-* and *bedh-*, however, *men-*, *med-*, and *meth-* come to be predicted as well. Furthermore, stems that are reported only in Pāṇini, such as *jer-* (to $\sqrt{jar^i}$ 'grow old') and *redh-* (to $\sqrt{rādh}$ 'harm'), come to be predicted as well. One possible point of overgeneration is that a stem *jen-* (to $\sqrt{jan^i}$ 'generate') is predicted, though such a stem is never attested or reported; given the high token frequency of the weak stem *jajñ-* (69× RV), which continues to be predicted until the third generation of the learning simulation, *jajñ-* could easily have persisted through lexical storage and retrieval.

In proceeding from a considerable set of forms possessed of a perfect weak stem of the form $C_1eC_2$- for purely phonological reasons, the path to all of such forms attested in the history of Sanskrit finally becomes open to us.

### 8.5.3 Zukoff 2015: The Poorly-Cued Repetition Principle

Zukoff (2015) has more recently offered a reinterpretation of his work on Greek reduplication, which he claims could also capture the existence and seeming analogical spread of $C_1eC_2$- forms in Vedic as a largely phonological development. In this work, Zukoff eschews the use of a constraint that makes reference to the domain of the syllable, but instead derives many of the same effects obtainable through the use of OCP-$\sigma$ with another constraint, Poorly-Cued Repetition:

(60) Poorly-Cued Repetition (PCR):
Assign a violation mark * to any $C_\alpha VC_\alpha$ sequence where the second $C_\alpha$ lacks the **requisite cues** to its presence.

In general, an obstruent-to-sonorant transition (TR or SR sequence) will be well-cued (by burst or frication noise plus and a rise in intensity), whereas stop-to-stop (TT sequence) or fricative-to-stop transitions (ST sequence) are more poorly cued (by either a burst or frication noise alone). In Zukoff's interpretation, it is precisely because ST is relatively poorly cued that Sanskrit roots such as $\sqrt{stambh}$ 'prop' form a perfect stem *tastambh-* rather than

$^{X}$*sastambh-*, in order to avoid a violation of the PCR. Zukoff follows Sandell 2013 in sustaining that forms like Ved. *pec-* and *bhej-* arose in PIIr. from the sheer phonotactic badness of the labial+palatal clusters. However, rather than attributing stems such as *sep-* and *pet-* to the extension of a morphological rule sensitive to phonological structure, he sees such forms as the tightening of restrictions on requisite cues in Sanskrit with respect to PIE and PIIr. While PIE/PIIr. would have permitted a *[sespr̥], because the fricative-stop transition would have been regarded as sufficiently well-cued (cf. Lat. *sistō*, Gk. ἵστημι [hístɛ:mi] 'I stand'; cf. Byrd 2015: Section 3.3.8 on reconstructing $C_1$ copy in PIE reduplicants), Sanskrit does not; since a candidate $^{X}$*paspur* is excluded by high-ranked Linearity constraints, *sepur* steps in as an alternative stem form provided by the morphology (i.e., by violating a Use-Reduplication constraint, like that in (51) above).

Zukoff's PCR constraint, at a practical level, replicates the effects of OCP-$\sigma$. A PIIr. version of the PCR will create the same set of surface perfect weak stems of the form *[$C_1$ə:$C_2$-] needed to feed into the successful model used in 8.5.2. Crucially, that morphological model is necessary to determine the set of $C_1eC_2$- stems that the morphological component of the grammar can provide to the phonology in the event of PCR and Linearity violations. This is clear for two reasons. First, as soon as PCR requirements became more strict in Sanskrit, we would expect for all $C_1eC_2$- forms to emerge at the same time when reduplication with the stop in roots like $\sqrt{stambh}$ became the rule: we would expect to find *pet-* already in the Family Books of the RV – the alternative stem *pet-* must have been made available by morphology later than Sanskrit's tightening of requisite cue restrictions. Second, the fact that forms such as the reduplicated aorist *apaptat* 'flew' maintain a PCR violating form through the history of Vedic shows that the morphology never provides an alternative stem *pet-* in such a case, simply because it is not a perfect.

The remaining problem for the PCR account would be to motivate the changes in what the phonology of Sanskrit takes to be a "requisite cue" for the presence of a segment. A crucial component, regardless of whether one operates with a PCR constraint or OCP-$\sigma$ constraint, is that $C_1eC_2$- allomorphs in Vedic be generated through phonologically sensitive morphological rules.

## 8.6 Summary and Conclusion

In this chapter, I have given a complete and detailed account of the emergence of the Vedic perfect weak stems like *sed-* (to $\sqrt{sad}$ 'sit') and *pec-* (to $\sqrt{pac}$ 'cook') that do not exhibit a clear trace of the expected reduplication. Under 8.2, I first discussed why the creation of novel $C_1eC_2$- forms cannot be meaningfully explored in terms of mere measures of productivity: because such forms are phonologically conditioned, it is necessary to treat them within a model that makes reference to phonological factors, namely, Minimal Generalization Learning. Under 8.3, I argued that the existing proposals concerning the $C_1eC_2$- perfects were entirely inadequate; an attempt to enlarge the set of phonologically explicable forms by reference to phonotactic constraints proved helpful, but still insufficient to the task. Finally, a set of phonologically grounded $C_1eC_2$- forms large enough to generate those that seem to

require a morphological basis could be obtained by reference to the OBLIGATORY CONTOUR PRINCIPLE.

This final analysis demonstrates that unconstrained usage of analogy as an explanatory mechanism may hide the true phonological generalizations and historical changes. In this particular case, by insisting upon a tightly constrained model of analogy, we find yet further support for the effects of OCP-$\sigma$ – it is only once we accept the operation of OCP-$\sigma$ or similar constraint in diachronic stages preceding Vedic that we are in a position to satisfactorily account for those few forms that, in the history of Vedic, genuinely require an analogical account. How much of what we believe about the historical phonology of old Indo-European languages (in their descent from their proto-languages) might be incomplete or incorrect because we have too eagerly shoe-horned forms into morphological (pseudo-)explanations, and thereby rendered the phonological truth invisible to ourselves?

### 8.6.1 Brief Excursus: *sed-* vs. *hazd-*: PIIr. *[səzd-] or *[səːd-]?

At first glance, the existence of the Avestan perfect stem *hazd-* (1×, in a 3.sg.perf.opt.act. *hazdiiāṭ*, Y. 65.5) would appear to point to the reconstruction of a PIIr. perfect stem *[səzd-]. If syllabically parsed as *[səz.dV-], the form would violate the strong version of OCP-$\sigma$ as given in (14) above. Whether *[zd-] was a licit onset in PIE (Byrd 2015: Appendix A does not reconstruct any certain instances of word-initial PIE *[zd-]) is not certain, though an OCP-$\sigma$ violation would have been inevitable in some paradigmatic forms, e.g., 1.pl. /sa-sd-má/. Under a PCR interpretation, given that *[sast-] likely did not violate the PCR in PIIr. (cf. Av. 3.pl.perf. *vi-šastarə* to $\sqrt{stā}$ 'stand'), then *[sezd-] ought to have been licit as well, unless the sequence of voiced fricative+stop is to be considered more poorly cued than the sequence of voiceless fricative+stop.

However, it is worth comparing the reduplicated present to $\sqrt{sad}$, Ved. 3.sg. *sī́dati*, Av. *hiδati*. While *[sížḍə-] would have been expected from /si-sd-ə-/, *[siždə-] patently cannot be the ancestor form of Vedic *sī́dati*, for which rather $^{\text{X}}$*sī́ḍati*, with retroflex *ḍ*, would be expected.[34] Vedic and Avestan then suggest that the PIIr. present stem was *[sīdə-], on account of OCP-$\sigma$. This conclusion would entail that the perfect weak stem */sə-sd-/ surface as *[səːd-].

What, then to make of Av. *hazd-*? First, even if one considers /h/ an independent phoneme of Avestan (despite the fact that /h/ and /s/ are largely in complementary distribution, and I find no minimal pairs), it is clear that /h/ is illicit preceding obstruents. A simple phonological rule, /h/ → [s] / _[-son], will ensure that a UR for the perfect stem to $\sqrt{had}$, /ha-hd-/, surfaces as Av. *hazd-*. The Avestan weak stem *hazd-* can thus be a synchronically generated perfect stem, and gives no necessary indication of what the PIIr. perfect stem may have been.

[34] To obtain Vedic *sī́dati* through a sporadic dissimilation in PIIr. (so Rix et al. 2001: 513–4, following Klingenschmitt 1982: 129) is *ad hoc*.

# Summary and Conclusions

At the outset of the Introduction, I raised the question of how "productivity" in general, and morphological productivity in particular, is to be grasped. From this core question sprung Part I; there, Chapters 1–3 served mainly to advance three claims:

- morphological productivity is most sensibly discussed and concretely grasped when assessed quantitatively, and a quantitative measure should, in principle, permit one to evaluate changes in productivity diachronically (Chapter 1).
- reliable methods for the measurement of morphological productivity based on corpus frequencies are available from the works of Baayen (in particular, the ratio of hapax legomena to tokens in a morphological category, $\mathscr{P}$), while research by Albright and Hayes furnishes a tool (the Minimal Generalization Learner) for the precise and probabilistic assessment of formal analogies as a kind of morphological change (Chapter 2).
- a large body of psycholinguistic research supports many assumptions underlying Baayen's measures of productivity and the Albright/Hayes approach to formal analogy; much of that research furthermore points to differences in morphological processing that are plausibly accounted for by "parsed" versus "holistic" models of lexical access (Chapter 3).

Part II, in turn, showed the successful practical application of those methods and tools to a diverse of data.

- Chapter 5 found that the productivity of the major categories of aorist (athematic, thematic, and sigmatic) in Ancient Greek, examining both Homer and the New Testament, could be quantitatively assessed without difficulty. Furthermore, the results obtained using Baayenian methods were found to accord with the intuitions of specialists of Ancient Greek. A simulation using the Minimal Generalization Learner based on the Homeric data further substantiated the productivity statistics. With the data from the New Testament, I illustrated how to directly compare productivity measures from two different corpora from two different historical periods.
- Chapter 6 similarly served to show the basic application of productivity measurement on Vedic data, again examining aorist categories. Just as results from Homer and the New Testament could be compared, so too was comparison between the results obtained from the *Ṛgveda* and Homer – this procedure might be the first concrete attempt at "comparative morphological productivity".
- Confident in the essential validity of corpus-based measures of productivity, Chapter 7 turned to a more slippery problem: the relation between morphological structure and word-level prosody in Greek and Sanskrit. I there first developed working analyses of accent assignment in both languages, influenced by but substantially diverging from

earlier work by Kiparsky. Attention to the non-productivity of particular derivational categories, and the frequencies of individual morphemes, appears to partially account for otherwise unaccountable variation in accentuation in lexemes belonging to the same morphological category.

- Chapter 8, finally, represents an attempt at a long-standing concern in the historical morphological of Sanskrit: the origin and analogical extension of weak stems of the perfect having the form $C_1eC_2$-. The objective was to arrive at a sufficient set of original forms with such a shape as to trigger further analogical spread. This condition was most readily obtained by accepting the existence of a phonotactic constraint, OCP-$\sigma$, which heavily penalized syllables containing identical or highly similar segments, in Proto-Indo-Iranian.

In the main, I believe that Chapters 5 and 6 together demonstrated that quantitative corpus-based measures of productivity can, by and large, be reliably extracted from corpora of the oldest Indo-European languages. This result is unsurprising: despite the many problems and peculiarities of the *Ṛgveda* and the Homeric epics as textual entities, they remain, fundamentally, samples of natural human languages, and hence exhibit the expected statistical properties thereof. More importantly, it appears to be the case that corpora of only $\sim$ 199000 (Homer) and $\sim$ 170000 tokens (*Ṛgveda*) are *not* too small to furnish statistically adequate data. These two studies also provide some baseline measurements for where unproductive and non-productive zones, quantitatively speaking, lie in the morphological categories attested in those corpora. However, what the systematic quantitative study of morphology, syntax, and semantics in older Indo-European languages truly requires, above all else, are more deeply tagged corpora. Indeed, the extent to which those factors interact in determining the behavior of certain phenomena will not be possible to exploit in more than a superficial fashion (or without painstaking and prohibitively time-consuming manual work) until corpora with such detailed information become available.

Chapters 7 and 8, for their part, each makes a theoretical contribution to the study of morphological change. While not fully predictive in itself, results in Chapter 7 suggest that morphological productivity, as measurable by the statistic $\mathscr{P}$, appears to be an important indicator of morphological structure. In languages with morphologically sensitive prosodic systems, like Greek and Vedic, low productivity and poor parsability may find themselves reflected in word prosody. My examination of Greek and Vedic accentuation further argued that many peculiarities in their accentual system are explicable by reference to morphological headedness (as per Revithiadou 1999); the absolute validity of the accentual grammars developed awaits, however, further empirical testing on large data sets. The study of a formal analogy in Sanskrit verbs in Chapter 8, meanwhile, arrived to a striking conclusion: insofar as analogies are to be constrained and modeled at all, not all need occur in the sweep of a single generation. Rather, the case of Sanskrit perfect weak stems suggests that analogies may "cascade" over the course of several generations: new members are gradually absorbed by a pattern, bit by bit, as a morphophonological mapping is able to grow in scope and increase in reliability.

The implicit and explicit goals set forth nearly 250 pages previously, from my vantage, appear fulfilled: we have tools that permit us to discuss productivity in very concrete terms. Furthermore, I have uncovered some phenomena that necessitated the application of those methods in order to arrive at a satisfying, or even approximative, solution. Let us now find further problems to attack similarly.

# Bibliography


Ackermann, Katsiaryna. 2015. *Die Vorgeschichte des slavischen Aoristsystems*. Leiden: Brill.

Agazzi, Pierangelo, and Massimo Vilardo. 2002. *Ἑλληνιστί. Grammatica della lingua greca*. Bologna: Zanichelli.

Albright, Adam. 2002a. Islands of Reliability for Regular Morphology: Evidence from Italian. *Language* 78.684–709.

———. 2002b. The Identification of Bases in Morphological Paradigms. Ph.D. diss., University of California, Los Angeles.

———. 2005. The Morphological Basis of Paradigm Leveling. In Laura Downing, Tracy Alan Hall and Renate Raffelsiefen (eds.), *Paradigms in Phonological Theory*, 17–43. Oxford: Oxford University Press.

———. 2008a. Explaining Universal Tendencies and Language Particulars in Analogical Change. In Jeff Good (ed.), *Language Universals and Language Change*, 144–80. Oxford: Oxford University Press.

———. 2008b. Modeling Analogy as Probabilistic Grammar. Manuscript: http://web.mit.edu/albright/.

———. 2010. Base-driven Leveling in Yiddish Verb Paradigms. *Natural Language and Linguistic Theory* 28.475–537.

———. 2012. Modeling Morphological Productivity with the Minimal Generalization Learner. URL `http://www2.uni-siegen.de/~engspra/draem/Albright-MGL-Tutorial.pdf`, Slides, Data-Rich Approaches to English Morphology, July 4–6.

Albright, Adam, and Bruce Hayes. 1999. An Automated Learner for Phonology and Morphology. Manuscript, UCLA.

———. 2003. Rules vs. Analogy in English Past Tenses: A Computational/Experimental Study. *Cognition* 90.119–61.

———. 2006. Modeling Productivity with the Gradual Learning Algorithm: The Problem of Accidentally Exceptionless Generalizations. In Gisbert Fanselow, Caroline Féry, Ralf Vogel and Matthias Schlesewsky (eds.), *Gradience in Grammar: Generative Perspectives*, 185–204. Oxford: Oxford University Press.

Alderete, John. 2001. *Morphologically Governed Accent in Optimality Theory*. London: Routledge.

Arnold, E. Vernon. 1905. *Vedic Metre in its Historical Development*. Cambridge: Cambridge University Press.

Aronoff, Mark. 1976. *Word Formation in Generative Grammar*. Cambridge, MA: MIT Press.

———. 1980. The Relevance of Productivity in a Synchronic Description of Word-Formation. In Jacek Fisiak (ed.), *Historical Morphology*, 71–2. The Hague: Mouton de Gruyter.

———. 2007. In the Beginning was the Word. *Language* 83.803–30.

Aronoff, Mark, and Frank Anshen. 1998. Morphology and the Lexicon: Lexicalization and Productivity. In A. Spencer and Arnold M. Zwicky (eds.), *The Handbook of Morphology*, 237–48. Oxford: Blackwell.

Aslin, Richard N. 1995. Infants' Detection of the Sound Patterns of Words in Fluent Speech. *Cognitive Psychology* 29.1–23.

Attneave, Fred. 1953. Psychological Probability as a Function of Experienced Frequency. *Journal of Experimental Psychology* 46.81–6.

Aufrecht, Theodor (ed.). 1877. *Die Hymnen des Rigveda*, 2nd edn. Bonn: A. Marcus.

Avery, John. 1880. Contributions to the History of Verb-Inflection in Sanskrit. *Journal of the American Oriental Society* 10.219–324.

Baayen, R. Harald. 1989. A Corpus-Based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistic Interpretation. Ph.D. diss., Vrije Universiteit Amsterdam.

———. 1992. Quantitative Aspects of Morphological Productivity. In Geert E. Booij and J. van Marle (eds.), *Yearbook of Morphology 1991*, 109–49. Dordrecht: Kluwer.

———. 1993. On Frequency, Transparency, and Productivity. In Geert E. Booij and Jaap van Marle (eds.), *Yearbook of Morphology 1992*, 181–208. Dordrecht: Kluwer.

———. 2001. *Word Frequency Distributions*. Dordrecht: Kluwer.

———. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

———. 2009. Corpus Linguistics in Morphology: Morphological Productivity. In Anke Lüdeling and M. Kyto (eds.), *Corpus Linguistics: An International Handbook*, 900–19. Mouton de Gruyter.

———. 2010. Demythologizing the Word Frequency Effect: A Discriminative Learning Perspective. *The Mental Lexicon* 5.436–61.

———. 2013 [2007]. languageR. URL `http://cran.r-project.org/web/packages/languageR/index.html`.

Baayen, R. Harald, Ton Dijkstra, and Robert Schreuder. 1997. Singulars and Plurals in Dutch: Evidence for a Parallel Dual Route Model. *Journal of Memory and Language* 37.94–117.

Baayen, R. Harald, and Rochelle Lieber. 1991. Productivity and English Derivation: A Corpus Based Study. *Linguistics* 29.801–43.

Baayen, R. Harald, Fermín Moscoso del Prado Martín, Robert Schreuder, and Lee Wurm. 2003. When Word Frequencies do not Regress towards the Mean. In R. Harald Baayen and Robert Schreuder (eds.), *Morphological Structure in Language Processing*, 463–84. Berlin: de Gruyter.

Baayen, R. Harald, and Antoinette Renouf. 1996. Chronicling the Times: Productive Lexical Innovations in an English Newspaper. *Language* 72.69–96.

Baayen, R. Harald, and Robert Schreuder. 1995. Modeling Morphological Processing. In Laurie Beth Feldman (ed.), *Morphological Aspects of Language Processing*, 131–54. New York: Psychology Press.

———. 1999. War and Peace: Morphemes and Full Forms in a Noninteractive Activation Parallel Dual-Route Model. *Brain and Language* 68.27–32.

———. 2000. Towards a Psycholinguistic Computational Model for Morphological Parsing. *Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Sciences* 358(1769, Computers, Language, and Speech: Formal Theories and Statistical Data).1281–93.

Baayen, R. Harald, and Robert Schreuder (eds.). 2003. *Morphological Structure in Language Processing*. Berlin: Mouton de Gruyter.

Baroni, Marco, and Stefan Evert. 2005. Testing the Extrapolation Quality of Word Frequency Models. In P. Danielsson and M. Wagenmakers (eds.), *Proceedings of Corpus Linguistics 2005, Proceedings from the Corpus Linguistics Conference Series*, vol. 1, 1–18.

Barth, Danielle, and Vsevolod Kapatsinski. 2015. A Multimodel Inference Approach to Categorical Variant Choice: Construction Priming and Frequency Effects on the Choice between Full and Contracted Forms of *am*, *are* and *is*. *Corpus Linguistics and Linguistic Theory* 11.1–58.

Bartholomae, Christian. 1885. Die altindischen *ē*-formen im schwachen perfekt. *Zeitschrift für vergleichende Sprachforschung* 27.337–66.

———. 1904. *Altiranisches Wörterbuch*. Strassburg: Trübner.

Bauer, Laurie. 1983. *English Word Formation*. Cambridge: Cambridge University Press.

———. 2001. *Morphological Productivity*. Cambridge: Cambridge University Press.

Beard, R. 1977. On the Extent and Nature of Irregularity in the Lexicon. *Lingua* 42.305–41.

Becker, Thomas. 1990. *Analogie und Morphologische Theorie*. München: Wilhelm Fink.

———. 1993. Morphologische Ersetzungsbildungen im Deutschen. *Zeitschrift für Sprachwissenschaft* 12.185–217.

Beckwith, Miles C. 1994. Greek ηὖρον, Laryngeal Loss and the Greek Reduplicated Aorist. *Glotta* 72.24–30.

———. 1996. The Greek Reduplicated Aorist. Ph.D. diss., Yale University.

Beekes, Robert Stephen Paul. 1969. *The Development of the Proto-Indo-European Laryngeals in Greek*. Den Haag: Mouton de Gruyter.

Benveniste, Émile. 1935. *Les infinitifs avestiques*. Paris: Adrien Maisonneuve.

Berko, Jean. 1958. The Child's Learning of English Morphology. *Word* 14.150–77.

Bertram, Raymond, Matti Laine, R. Harald Baayen, Robert Schreuder, and Jukka Hyönä. 2000. Affixal Homonymy Triggers Full-Forms Storage, Even with Inflected Words, Even in a Morphologically Rich Language. *Cognition* 74.B13–25.

Bien, Hiedrun, R. Harald Baayen, and Willem J. M. Levelt. 2011. Frequency Effects in the Production of Dutch Deverbal Adjectives and Inflected Verbs. *Language and Cognitive Processes* 27.683–715.

Bloomfield, Leonard. 1931 [1984]. *Language*. Chicago: University of Chicago Press.

Bod, Rens, Jennifer Hay, and Stefanie Jannedy (eds.). 2003. *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Boersma, Paul, and David Weenink. 1992–2014. Praat. Doing Phonetics by Computer. Software. URL `www.praat.org`.

von Böhtlingk, Otto. 1887 [1964]. *Pāṇini's Grammatik*. Hildesheim: Georg Olms.

Bolinger, Dwight. 1948. On Defining the Morpheme. *Word* 4.18–23.

Bolozky, Shmuel. 1999. *Measuring Productivity in Word Formation*. Leiden: Brill.

Bonet, Eulàlia, Maria-Rosa Lloret, and Joan Mascaró (eds.). Forthcoming. *Understanding Allomorphy. Perspectives from Optimality Theory*. London: Equinox.

Booij, Geert E. 2007. *The Grammar of Words*. Oxford: Oxford University Press.

———. 2010. *Construction Morphology*. Oxford: Oxford University Press.

Borer, Hagit. Forthcoming. The Syntactic Domain of Content. In J. Grinstead, J. Rothman and B. D. Schwartz (eds.), *Generative Linguistics and Acquisition: Studies in Honor of Nina M. Hyams*. Amsterdam: John Benjamins.

Bozzone, Chiara. 2014. The Origin of the Caland System and the Typology of Adjectives. Presentation. 33rd East Coast Indo-European Conference, Blacksburg, VA.

Branchaw, Sherrylyn Elizabeth. 2010. Survival of the Strongest: Strong Verbs in the History of English. Ph.D. diss., University of California, Los Angeles.

Bresnan, Joan, and Marilyn Ford. 2010. Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English. *Language* 86.186–213.

Březina, Vaclav. 2005. The Development of the Prefixes *un-* and *in-* in Early Modern English with Special Regard to Sociolinguistic Background. MA thesis, Charles University.

Buck, Carl Darling. 1955. *The Greek Dialects*. Chicago: University of Chicago.

Buck, Carl Darling, and Walter Petersen. 1949. *A Reverse Index of Greek Nouns and Adjectives*. Chicago: University of Chicago Press.

Burani, Cristina, and Anna M. Thornton. 2003. The Interplay of Root, Suffix and Whole-Word Frequency in Processing Derived Words. In R. Harald Baayen and Robert Schreuder (eds.), *Morphological Structure in Language Processing*, 157–208. Berlin: de Gruyter.

Burnett, Sarah A., and John M. Stevenson. 1979. Instructional Effects on Frequency Judgments. *American Journal of Psychology* 92.711–21.

Butterworth, Brian. 1983. Lexical Representation. In Brian Butterworth (ed.), *Language Production*, vol. 2, 257–94. New York: Academic Press.

Bybee, Joan. 2007. *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.

Bybee, Joan, and Dan Slobin. 1982. Rules and Schemas in the Development of the English Past Tense. *Language* 59.251–70.

Byrd, Andrew Miles. 2015. *The Indo-European Syllable*. Leiden: Brill.

Caland, Willem. 1892. Beiträge zur Kenntnis des Avesta, no. 19. *Zeitschrift für vergleichende Sprachforschung* 31.266–8.

Cannon, Garland. 1988. *Historical Change and English Word-Formation*. New York: Peter Lang.

Cardona, George. 1960. The Indo-European Thematic Aorists. Ph.D. diss., Yale University.

————. 1965. The Vedic Imperatives in *-si*. *Language* 41.1–18.

Cathcart, Chundra. 2012. Vedic Labial Dissimilation Revisited. In Stephanie W. Jamison, H. Craig Melchert and Brent Vine (eds.), *Proceedings of the 23rd Annual UCLA Indo-European Conference*, 29–45. Bremen: Hempen.

Chantraine, Pierre. 1958. *Grammaire homérique. Tome I. Phonétique et morphologie*, 3rd edn. Paris: Klincksieck.

Charnov, Eric L. 1976. Optimal Foraging: The Marginal Value Theorem. *Theoretical Population Biology* 9.129–36.

Chater, Nick, Joshua B. Tenenbaum, and Alan Yuille. 2006. Probabilistic Models of Cognition: Conceptual Foundations. *Trends in Cognitives Sciences* 10.287–91.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

————. 1970. Remarks on Nominalization. In Roderick A. Jacobs and Peters S. Rosenbaum (eds.), *Readings in Transformational Grammar*, 184–221. Waltham, MA: Ginn and Co.

————. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper & Row.

Clements, George N. 1990. The Role of the Sonority Cycle in Core Syllabification. In John Kingston and Mary Beckman (eds.), *Papers in Laboratory Phonology 1: Between the Grammar and Physics of Speech*, 283–333. Cambridge: Cambridge University Press.

Coetzee, Andries. 2014. Grammatical Change through Lexical Accumulation: Voicing Cooccurrence Restrictions in Afrikaans. *Language* 90.693–721.

Coetzee, Andries, and Shigeto Kawahara. 2013. Frequency Biases in Phonological Variation. *Natural Language and Linguistic Theory* 31.47–89.

Coppieters, René. 1987. Competence Differences between Native and Near-Native Speakers. *Language* 63.544–73.

Corbin, Danielle. 1980. Compétence lexicale et compétence syntaxique. *Modèles Linguistiques* II.2.52–138.

Cowie, Claire, and Christiane Dalton-Puffer. 2002. Diachronic Word-Formation and Studying Changes in Productivity over Time: Theoretical and Methodological Considerations. In Javier E. Díaz-Vera (ed.), *A Changing World of Words: Studies in English Historical Lexicography, Lexicology, and Semantics*, 410–37. Amsterdam: Rodopi.

Dąbrowska, Ewa. 2008. The Effects of Frequency and Neighborhood Density on Adult Speakers' Productivity with Polish Case Inflections: An Empirical Test of Usage-Based Approaches in Morphology. *Journal of Memory and Language* 58.931–51.

Daelemans, Walter, and Antal van den Bosch. 2005. *Memory-Based Language Processing*. Cambridge: Cambridge University Press.

Dahl, Eystein. 2010. *Time, Tense and Aspect in Early Vedic Grammar: Exploring Inflectional Semantics in the Rigveda*. Leiden: Brill.

———. 2013. Typological Change in Vedic: The Development of the Aorist from a Perfective Past to an Immediate Past. In Folke Josephson and Ingmar Söhrman (eds.), *Diachronic and Typological Perspectives on Verbs*, 261–98. Amsterdam: John Benjamins.

Dalton-Puffer, Christiane. 1996. *The French Influence on Middle English Morphology: A Corpus-Based Study of Derivation*. Berlin: Mouton de Gruyter.

Derksen, Rick. 2015. *Etymological Dictionary of the Baltic Inherited Lexicon*. Leiden: Brill.

Diependaele, Kevin, Jonathan Grainger, and Dominiek Sandra. 2012. Derivational Morphology and Skilled Reading: An Empirical Overview. In Michael J. Spivey, Ken McRae and Marc F. Joanisse (eds.), *The Cambridge Handbook of Psycholinguistics*, 311–32. Cambridge: Cambridge University Press.

Divjak, Dagmar, and Stefan Th. Gries (eds.). 2012. *Frequency Effects in Language Representation*. Berlin: de Gruyter.

Dixon, R. M. W. 2004. *The Jarawara Language of Southern Amazonia*. Oxford: Oxford University Press.

———. 2010. *Basic Linguistic Theory*. Oxford: Oxford University Press.

Dressler, Wolfgang. 1985. On the Predictiveness of Natural Morphology. *Journal of Linguistics* 21.321–37.

———. 1987. Word Formation as Part of Natural Morphology. In Wolfgang Dressler (ed.), *Leitmotifs in Natural Morphology*, 99–126. Amsterdam: John Benjamins.

———. 1999. On a Semiotic Theory of Preferences in Language. In *The Peirce Seminar Papers*, vol. 4, 389–415. New York: Berghahn Books.

Dybo, Vladimir A. 1968. Akcentologija i slovoobrazovanie v slavjanskom. In *Slavjanskoe jazykoznanie, VI Meždunarodnyj S'ezd Slavistov*, 148–224. Moskva: Nauka.

Eddington, David. 2000. Analogy and the Dual-Route Model of Morphology. *Lingua* 110.281–98.

Ellis, Nick C. 2012. What Can We Count in Language, and What Counts in Language Acquisition, Cognition, and Use? In Stefan Th. Gries and Dagmar Divjak (eds.), *Frequency Effects in Language Learning and Processing*, 7–34. Berlin: de Gruyter.

Ernestus, Miriam, and Harald Baayen. 2003. Predicting the Unpredictable: Interpreting Neutralized Segments in Dutch. *Language* 79.5–38.

Evert, Stefan. 2004. A Simple LNRE Model for Random Character Sequences. In G. Purnelle, C. Fairon and A. Dister (eds.), *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis*, 411–22. Louvain-la-Neuve: UCL.

Evert, Stefan, and Marco Baroni. 2008. zipfR: Lexical Statistics in R. URL `http://zipfr.r-forge.r-project.org/zipfR`.

Evert, Stefan, and Anke Lüdeling. 2001. Measuring Morphological Productivity: Is Automatic Preprocessing Sufficient? In P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds.), *Proceedings of the Corpus Linguistics 2001 Conference*, 167–75. Lancaster: UCREL.

Forbes, Kathleen. 1958. Medial Intervocalic -ρσ-, -λσ- in Greek. *Glotta* 36.235–72.

Ford, Michael A., William D. Marslen-Wilson, and Matthew H. Davis. 2003. Morphology and Frequency: Contrasting Methodologies. In R. Harald Baayen and Robert Schreuder (eds.), *Morphological Structure in Language Processing*, 89–124. Berlin: Mouton de Gruyter.

Fortson, Benjamin W. 2010. *Indo-European Language and Culture*, 2nd edn. Oxford: Wiley-Blackwell.

Franceschini, Marco. 2007. *An Updated Vedic Concordance: Maurice Bloomfield's A Vedic Concordance with New Material Taken from Seven Vedic Texts*. Cambridge, MA: Harvard University Press.

Frauenfelder, Ulrich H., and Robert Schreuder. 1992. Constraining Psycholinguistic Models of Morphological Processing and Representation: The Role of Productivity. In Geert E. Booij and Jaap van Marle (eds.), *Yearbook of Morphology 1991*, 165–83. Dordrecht: Kluwer.

Frisch, Stefan A., Janet B. Pierrehumbert, and Michael B. Broe. 2004. Similarity Avoidance and the OCP. *Natural Language and Linguistic Theory* 22.179–228.

Frost, R., T. Kugler, A. Deutsch, and K. I. Forster. 2005. Orthographic Structure Versus Morphological Structure: Principles of Lexical Organization in a Given Language. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31.1293–1326.

Gaeta, Livio. 2007. On the Double Nature of Productivity in Inflectional Morphology. *Morphology* 17.181–205.

Gaeta, Livio, and Davide Ricca. 2006. Productivity in Italian Word Formation: A Variable-Corpus Approach. *Linguistics* 44.57–89.

Gallistel, Charles R. 1990. *The Organization of Learning*. Cambridge, MA: Bradford Books/MIT Press.

Gardani, Francesco. 2013. *Dynamics of Morphological Productivity. The Evolution of Noun Classes from Latin to Italian*. Leiden: Brill.

Garnier, Roman. 2010. *Tum mihi prīma genās*: Phraséologie et étymologie du Latin *pūbēs*. *Historische Sprachforschung* 123.181–211.

Geldner, Karl Friedrich. 1951. *Der Rig-Veda. Aus dem Sanskrit ins Deutsche übersetzt und mit einem laufendem Kommentar versehen*. Cambridge, MA: Harvard University Press.

Gernsbacher, Morton A. 1984. Resolving 20 Years of Inconsistent Interactions between Lexical Familiarity and Orthography, Concreteness, and Polysemy. *Journal of Experimental Psychology* 113.256–81.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT Constraint Rankings Using a Maximum Entropy Model. In Jennifer Spenader, Anders Eriksson and Östen Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, 111–20. Stockholm University Department of Linguistics.

Golston, Chris. 1990. Floating H (and L*) Tones in Ancient Greek. In J. Meyers and P.E. Pérez (eds.), *Arizona Phonology Conference, iii*, 66–82. Tucson, Arizona: University of Arizona Linguistics Department.

Gómez, Rebecca L., and LouAnn Gerken. 1999. Artificial Grammar Learning by 1-year-olds Leads to Specific and Abstract Knowledge. *Cognition* 70.109–35.

———. 2000. Infant Artificial Language Learning and Language Acquisition. *Trends in Cognitive Sciences* 4.178–86.

Good, Irving John. 1953. The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika* 40.237–64.

Gotō, Toshifumi. 1987. *Die "I. Präsensklasse" im Vedischen. Untersuchung der vollstufigen thematischen Wurzelpräsentia*. Wien: Österreichischen Akademie der Wissenschaften.

Grassmann, Hermann. 1872 [1976]. *Wörterbuch zum Rig-Veda*, 5th edn. Wiesbaden: Harrassowitz.

Gries, Stephan Th. 2008. Dispersions and Adjusted Frequencies in Corpora. *International Journal of Corpus Linguistics* 13.403–37.

Gunkel, Dieter. 2010. Studies in Greek and Vedic Prosody, Morphology, and Meter. Ph.D. diss., University of California, Los Angeles.

———. 2013. Tonal Ochlophobia in Greek. Presentation. 32nd East Coast Indo-European Conference, Adam Mickiewicz University Poznań, 21–24 June.

Halle, Morris. 1973. Prolegomena to a Theory of Word Formation. *Linguistic Inquiry* 4.3–16.

Halle, Morris, and Alec Marantz. 1993. Distributed Morphology and the Pieces of Inflection. In *The View from Building 20*, 111–76. Cambridge, MA: MIT Press.

———. 1994. Some Key Features of Distributed Morphology. In *Papers on Phonology and Morphology, MIT Working Papers in Linguistcs*, vol. 21, 275–88. MIT Working Papers in Linguistics.

Halle, Morris, and Jean-Roger Vergnaud. 1987. *An Essay on Stress*. Cambridge, MA: MIT Press.

Harper, D. G. C. 1982. Competitive Foraging in Mallards: Ideal Free Ducks. *Animal Behaviour* 30.575–84.

Harðarson, Jón Axel. 1993. *Studien zum urindogermanischen Wurzelaorist und dessen Vertretung im Indoiranischen und Griechischen*. Innsbruck: Innsbrucker Beiträge zur Sprachwissenschaft.

Hasher, Lynn, David Goldstein, and Thomas Toppino. 1977. Frequency and the Conference of Referential Validity. *Journal of Verbal Learning and Behavior* 16.107–12.

Hasher, Lynn, and Rose T. Zacks. 1979. Automatic and Effortful Processes in Memory. *Journal of Experimental Psychology: General* 108.356–88.

———. 1984. Automatic Processing of Fundamental Information: The Case of Frequency of Occurrence. *American Psychologist* 39.1372–88.

Haslam, Michael. 1997 [2011]. Homeric Papyri and the Transmission of the Text. In Ian Morris and Barry Powell (eds.), *A New Companion to Homer*, 55–100. Leiden: Brill.

Hay, Jennifer. 2001. Lexical Frequency in Morphology: Is Everything Relative? *Linguistics* 39.1041–70.

———. 2003. *Causes and Consequences of Word Structure*. New York: Taylor & Francis.

———. 2007. The Phonetics of 'un'. In Judith Munat (ed.), *Lexical Creativity, Texts and Contexts*, 39–57. Amsterdam: John Benjamins.

Hay, Jennifer, and R. Harald Baayen. 2002. Parsing and Productivity. In Geert E. Booij and Jaap van Marle (eds.), *Yearbook of Morphology 2001*, 203–35. Dordrecht: Kluwer.

———. 2003. Phonotactics, Parsing, and Productivity. *Italian Journal of Linguistics* 15.99–130.

Hayes, Bruce. 1989. Compensatory Lengthening in Moraic Phonology. *Linguistic Inquiry* 20.253–306.

———. 2004. Phonological Acquisition in Optimality Theory: The Early Stages. In René Kager, Joe Pater and Wim Zonneveld (eds.), *Constraints in Phonological Acquisition*, 158–203. Cambridge: Cambridge University Press.

Hayes, Bruce, and Colin Wilson. 2008. A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *Linguistic Inquiry* 39.379–440.

Hayes, Bruce, Kie Zuraw, Péter Siptár, and Zsuzsa Cziráky Londe. 2009. Natural and Unnatural Constraints in Hungarian Vowel Harmony. *Language* 85.822–63.

Herrnstein, R. J. 1961. Relative and Absolute Strength of Response as a Function of Frequency of Reinforcement. *Journal of the Experimental Analysis of Behavior* 4.267–72.

Hill, Eugen. 2012. Hidden Sound Laws in the Inflectional Morphology of Proto-Indo-European. A Phonological Account of the Primary First Singular of Thematic Verbs and the Instrumental Case of Thematic Nouns. In Benedicte Nielsen Whitehead, Thomas Olander, Birgit A. Olsen and Jens E. Rasmussen (eds.), *The Sound of Indo-European. Phonetics, Phonemics, and Morphophonemics.* Copenhagen: Museum Tusculanum.

Hill, Jane H., and Kenneth C. Hill. 1968. Stress in the Cupan (Uto-Aztecan) Languages. *International Journal of American Linguistics* 34.233–41.

Hinge, George. 2006. *Die Sprache Alkmans*. Wiesbaden: Reichert.

Hintzman, Douglas. 1969. Apparent Frequency as a Function of Frequency and the Spacing of Repetitions. *Journal of Experimental Psychology* 80.139–45.

———. 1976. Repetition and Memory. In G. H. Bower (ed.), *The Psychology of Learning and Motivation*, vol. 10, 47–91. New York: Academic Press.

Hirt, Herman Alfred. 1900. *Der indogermanische Ablaut: vornehmlich in seinem Verhältnis zur Betonung*. Strassburg: Trübner.

Hock, Hans Heinrich. 1991. *Principles of Historical Linguistics*, 2nd edn. Berlin: Mouton de Gruyter.

Hoffmann, Karl. 1967. *Der Injunktiv im Veda*. Heidelberg: Carl Winter.

Hoffner, Harry A., and H. Craig Melchert. 2008. *A Grammar of the Hittite Language*. Winona Lake: Eisenbrauns.

Holmes, Michael W. (ed.). 2010. *The Greek New Testament*. Atlanta: Society of Biblical Literature.

Howes, D. H., and R. L. Solomon. 1951. Visual Duration Threshold as a Function of Word Probability. *Journal of Experimental Psychology* 41.401–10.

Illich-Svitych, Vladislav M. 1963. *Imennaja akcentuacija v baltijskom i slavjanskom*. Moscow: Institut Slavjanovedenija.

———. 1979. *Nominal Accentuation in Baltic and Slavic.* Cambridge, MA: MIT Press. Translation of *Imennaja akcentuacija v baltijskom i slavjanskom*, Moscow 1963, by Richard L. Leed and Ronald F. Feldstein.

Jackendoff, Ray. 1975. Morphological and Semantic Regularities in the Lexicon. *Language* 51.639–71.

———. 2002. What's in the Lexicon? In Sieb Nooteboom, Fred Weerman and Frank Wijnen (eds.), *Storage and Computation in the Language Faculty*, 23–60. Dordrecht: Kluwer.

Jamison, Stephanie. 1983. *Function and Form in the* -áya- *Formations of the Rig Veda and Atharva Veda*. Göttingen: Vandenhoeck & Ruprecht.

Jamison, Stephanie W., and Joel P. Brereton. 2014. *The Rigveda. The Earliest Religious Poetry of India*. Oxford: Oxford University Press.

Jasanoff, Jay H. 2008. The Accentual Type **vèdō*, **vedetí* and the Origin of Mobility in the Balto-Slavic Verb. *Baltistica* 43.339–79.

————. 2012. Long-vowel Preterites in Indo-European. In H. Craig Melchert (ed.), *The Indo-European Verb. Proceedings of the Conference of the Society for Indo-European Studies, Los Angeles 13-15 September 2010*, 127–35. Wiesbaden: Reichert.

Johnson, Keith. 2008. *Quantitative Methods in Linguistics*. Malden, MA: Blackwell.

de Jong, Nivja H., Robert Schreuder, and R. Harald Baayen. 2000. The Morphological Family Size Effect and Morphology. *Language and Cognitive Processes* 15.329–65.

————. 2003. Morphological Resonance in the Mental Lexicon. In R. Harald Baayen and Robert Schreuder (eds.), *Morphological Structure in Language Processing*, 65–88. Berlin: Mouton de Gruyter.

Jurafsky, Dan. 2003. Probabilistic Modeling in Psycholinguistic Comprehension and Production. In Rens Bod, Jennifer Hay and Stefanie Jannedy (eds.), *Probabilistic Linguistics*, 39–95. Cambridge, MA: MIT Press.

Kaiser, Stefan, Yasuko Ichikawa, Noriko Kobayashi, and Hilofumi Yamamoto. 2001 [2013]. *Japanese: A Comprehensive Grammar*, 2nd edn. New York: Routledge.

Kapatsinski, Vsevolod. 2013. Conspiring to Mean: Experimental and Computational Evidence for a Usage-Based Harmonic Approach to Morphophonology. *Language* 89.110–48.

Katz, Leonard, Karl Rexer, and Georgije Lukatela. 1991. The Processing of Inflected Words. *Psychological Research* 53.25–32.

Kawahara, Shigeto, and Shin-ichiro Sano. 2014. Identity Avoidance and Lyman's Law. *Lingua* 150.71–7.

Kelly, Michael H., and Susane Martin. 1994. Domain-General Abilities Applied to Domain-specific Tasks: Sensitivity to Probabilities in Perception, Cognition, and Language. *Lingua* 92.105–40.

Keuleers, Emmanuel. 2008. Memory-Based Learning of Inflectional Morphology. Ph.D. diss., Universiteit Antwerpen.

Keydana, Götz. 2014. Ablaut in indogermanischen Primärnomina: Die hysterokinetischen Stämme. In Norbert Oettinger and Thomas Steer (eds.), *Das Nomen im Indogermanischen. Morphologie, Substantiv versus Adjektiv, Kollektivum. Akten der Arbeitstagung der Indogermanischen Gesellschaft, Erlangen 14-16. September 2011*, 113–28. Wiesbaden: Reichert.

Khalmadze, Estate. 1987. *The Statistical Analysis of a Large Number of Rare Events*. Tech. Rep. MS–R8804, Department of Mathematical Statistics, Center for Mathematics and Computer Science, Amsterdam.

Kim, Ronald. 2002. Topics in the Reconstruction and Development of Indo-European Accent. Ph.D. diss., University of Pennsylvania.

Kiparsky, Paul. 1973. The Inflectional Accent in Indo-European. *Language* 49.794–849.

———. 1982a. *Explanation in Phonology*. Dordrecht: Foris.

———. 1982b. From Cyclic Phonology to Lexical Phonology. In Harry van der Hulst and Norval Smith (eds.), *The Structure of Phonological Representations*. Dordrecht: Foris.

———. 2003. Accent, Syllable Structure, and Morphology in Ancient Greek. In Elizabeth Mela Athanasopoulou (ed.), *Selected Papers from the 15th International Symposium on Theoretical and Applied Linguistics*, 81–106. Thessaloniki: Aristotle University of Thessaloniki.

———. 2010. Compositional vs. Paradigmatic Approaches to Accent and Ablaut. In Stephanie W. Jamison, H. Craig Melchert and Brent Vine (eds.), *Proceedings of the 21st UCLA Indo-European Conference*, 137–82. Bremen: Hempen.

———. Forthcoming. Accent and Ablaut. In Michael Weiss and Andrew Garrett (eds.), *Handbook of Indo-European Studies*. Oxford: Oxford University Press.

Klingenschmitt, Gert. 1982. *Das altarmenische Verbum*. Wiesbaden: Reichert.

Knobl, Werner. 2009. A Surplus of Meaning: The Intent of Irregularity in Vedic Poetry. Ph.D. diss., Universiteit Leiden.

Krämer, Martin. 2012. *Underlying Representations*. Cambridge: Cambridge University Press.

Krisch, Thomas. 1996. *Zur Genese und Funktion der altindischen Perfekta mit langem Reduplikationsvokal*. Innsbruck: Innsbrucker Beiträge zur Sprachwissenschaft.

Krott, Andrea, Robert Schreuder, R. Harald Baayen, and Wolfgang Dressler. 2007. Analogical Effects on Linking Elements in German Compounds. *Language and Cognitive Processes* 22.25–57.

Kubozono, Haruo. 1995. Constraint Interaction in Japanese Phonology: Evidence from Compound Accent. In Rachel Walker, Ove Lorentz and Haruo Kubozono (eds.), *Phonology at Santa Cruz*, vol. 4, 21–38. University of California, Santa Cruz.

Kučera, H., and W. N. Francis. 1967. *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Kulikov, Leonid. 2005. Reduplication in the Vedic Verb: Indo-European Inheritance, Analogy, and Iconicity. In Bernard Hurch (ed.), *Studies on Reduplication*, 431–54. Berlin: Mouton de Gruyter.

Kümmel, Martin Joachim. 2000. *Das Perfekt im Indoiranischen*. Wiesbaden: Reichert.

————. 2005. Vedisch *tand-* und ein neues indoiranisches Lautgesetz. In Günter Schweiger (ed.), *Indogermanica. Festschrift für Gert Klingenschmitt*, 321–32. Regensberg: Taimering.

van de Laar, Henri M. F. M. 2000. *Description of the Greek Individual Verbal Systems*. Amsterdam: Rodopi.

Labov, William. 1969. Contraction, Deletion, and Inherent Variability of the English Copula. *Language* 45.715–62.

de Lamberterie, Charles. 2009. En hommage à Michel Lejeune: Mycénien *o-wo-we* et le nom de l'«oreille» en grec. In Frédérique Biville and Isabelle Boehm (eds.), *Autour de Michel Lejeune. Actes des Journées d'étude organisées à l'Université Lumière-Lyon 2 — Maison de l'Orient et de la Méditerranée. 2-3 février 2006*, 79–116. Lyon: Maison de l'Orient et de la Méditerranée.

Lantz, Brett. 2013. *Machine Learning with R*. Birmingham - Mumbai: PACKT Publishing.

Laudanna, Alessandro, Cristina Burani, and Antonella Cermele. 1994. Prefixes as Processing Units. *Language and Cognitive Processes* 9.295–316.

Lehrer, Adrienne. 2007. Blendalicious. In Judith Munat (ed.), *Lexical Creativity, Texts and Contexts*, 115–33. Amsterdam: John Benjamins.

Lejeune, Michel. 1972. *Phonétique historique du mycénien et du grec ancien*. Paris: Klincksieck.

Leskien, August. 1891. *Die Bildung der Nomina im Litauischen. Abhandlung der Königlich Sächischen Gesellschaft der Wissenschaften*, vol. 12. Leipzig: S. Hirzel.

Leumann, Manu. 1962. Der altindische kausative Aorist *ajījanat*. In E. Bender (ed.), *Indological Studies in Honor of W. Norman Brown*, 152–59. New Haven: American Oriental Society.

Logos Bible Software. 2000–2014. Logos Bible Software. URL `www.logos.com`.

Lord, Albert Bates. 1960. *The Singer of Tales*. Cambridge, MA: Harvard University.

Lowe, John J. 2013. Aorist Participles in the Ṛgveda. Presentation. 25th Annual UCLA Indo-European Conference, 26 October.

Lubotsky, Alexander. 1998. *A Ṛ̊gVedic Word Concordance. American Oriental Series*, vol. 83. New Haven: American Oriental Society.

———. 2013. Dissimilatory Loss of *i* in Sanskrit. In Roman Sukač and Ondřej Šefčik (eds.), *The Sound of Indo-European 2. Papers on Indo-European Phonetics, Phonemics, and Morphophonemics*, 177–81. München: LINCOM.

Lundquist, Jesse. 2014. Vedic -tí- Abstracts and the Reconstruction of Proterokinetic *-tí- Stems in PIE. Presentation. 26th UCLA Indo-European Conference, October 24–25.

———. 2015. Vedic -tí- Abstracts: History and Prehistory. Presentation. Harvard GSAS Colloquium, February 25.

Macdonell, Arthur Anthony. 1910. *Vedic Grammar*. Strassburg: Trübner.

———. 1916 [1993]. *A Vedic Grammar for Students*. Delhi: Motilal Banarsidass.

Mandelbrot, Benoit. 1953. An Information Theory of the Statistical Structure of Language. In W. E. Jackson (ed.), *Communication Theory*, 503–12. New York: Academic Press.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marcus, Gary F., Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, and Fei Xu. 1992. *Overregularization in Language Acquisition*. *Monographs of the Society for Research in Child Development*, vol. 57. Chicago: University of Chicago Press.

van Marle, Jaap. 1985. *On the Paradigmatic Dimension of Morphological Productivity*. Dordrecht: Foris.

Marslen-Wilson, William, Lorraine K. Tyler, Rachelle Waksler, and Lianne Older. 1994. Morphology and Meaning in the English Mental Lexicon. *Psychological Review* 101.3–33.

Mascaró, Joan, Eulàlia Bonet, and Maria-Rosa Lloret. 2007. Allomorph Selection and Lexical Preferences: Two Case Studies. *Lingua* 117.903–27.

Maye, Jessica, Janet F. Werker, and LouAnn Gerken. 2002. Infant Sensitivity to Distribution Information Can Affect Phonetic Discrimination. *Cognition* 82.B101–11.

Mayerthaler, Willi. 1981. *Morphologische Natürlichkeit*. Wiesbaden: Athenaion.

Mayrhofer, Manfred. 1986–2001. *Etymologisches Wörterbuch des Altindoarischen*. Heidelberg: Carl Winter.

McCarthy, John, and Alan Prince. 1995. Faithfulness and Reduplicative Identity. In J. N. Beckman and Susan Urbanczyk (eds.), *Papers in Optimality Theory*, *University of Massachusetts Occasional Papers in Linguistics*, vol. 18, 249–384. University of Massachusetts, Amherst.

McClure, Scott Nathanael. 2011. A System for Inducing the Phonology and Inflectional Morphology of a Natural Language. Ph.D. diss., Yale University.

McPherson, Laura. 2014. Replacive Grammatical Tone in the Dogon Languages. Ph.D. diss., University of California, Los Angeles.

Meier-Brügger, Michael. 2003. *Indo-European Linguistics*. Berlin: Walter de Gruyter.

Meiser, Gerhard. 2003. *Veni Vidi Vici: Die Vorgeschichte des lateinischen Perfektsystems*. München: C. H. Beck.

Meissner, Torsten. 2006. *S-Stem Nouns and Adjectives in Greek and Proto-Indo-European: A Diachronic Study in Word Formation*. Oxford: Oxford University Press.

Melchert, H. Craig. 2014. "Narten formations" versus "Narten roots". *Indogermanische Forschungen* 119.251–8.

Mikheev, A. 1997. Automatic Rule Induction for Unknown-Word Guessing. *Computational Linguistics* 23.405–23.

Mohanen, K. 1986. *The Theory of Lexical Phonology*. Dordrecht: D. Reidel.

Monro, David Binning (ed.). 1902. *Iliad*. Oxford: Oxford University Press.

Moreton, Elliot. 2012. Structure and Substance in Artificial-Phonology Learning. Part I, Structure. Part II, Substance. *Language and Linguistics Compass* 6.686–701 and 702–18.

Mos, Maria. 2010. Complex Lexical Items. Ph.D. diss., Universiteit Tilburg.

Murray, A. T. (ed.). 1919. *Odyssey. With an English Translation by A. T. Murray*. Cambridge, MA: Harvard University Press.

Nagy, Gregory. 2004. *Homer's Text and Language*. Champaign, IL: University of Illinois Press.

Narten, Johanna. 1964. *Die sigmatischen Aoriste im Veda*. Wiesbaden: Reichert.

Neri, Sergio. 2011. Wetter. Etymologie und Lautgesetz. Ph.D. diss., Friedrich-Schiller Universität Jena.

Nevalainen, T., J. Keränen, M. Nevala, A. Nurmi, M. Palander-Collin, and H. Raumolin-Brunberg (eds.). 1998. *Corpus of Early English Correspondence*. Department of English, University of Helsinki. URL `http://www.helsinki.fi/varieng/domains/CEEC.html`.

Nosofsky, Robert M. 1990. Relations between Exemplar Similarity and Likelihood Models of Classification. *Journal of Mathematical Psychology* 34.393–418.

Noyer, Rolf. 1997. Attic Greek Accentuation and Intermediate Derivational Representations. In Iggy Roca (ed.), *Derivations and Constraints in Morphology*, 501–27. Oxford: Clarendon Press.

————. 2001. Clitic Sequences in Nunggubuyu and PF Convergence. *Natural Language and Linguistic Theory* 19.751–826.

Nussbaum, Alan. 1976. Caland's "Law" and the Caland System. Ph.D. diss., Harvard University.

————. 1987. Homeric ἐπελήκεον (Θ 379) and Related Forms. In Calvert Watkins (ed.), *Studies in Memory of Warren Cowgill (1929–1985). Papers from the Fourth East Coast Indo-European Conference, Cornell University, June 6–9, 1985*, 229–53. Berlin: de Gruyter.

OED. 2013. Oxford English Dictionary Online. URL `www.oed.com`.

Oettinger, Norbert. 1976. Der indogermanische Stativ. *Münchener Studien zur Sprachwissenschaft* 34.109–49.

Oldenberg, Hermann. 1909–12. *Ṛgveda. Textkritische und exegetische Noten*. Berlin: Weidmannsche Buchhandlung.

O'Neill, Eugene. 1942. The Localization of Metrical Word-Types in the Hexameter: Homer, Hesiod, and the Alexandrians. *Yale Classical Studies* 8.103–78.

Panagl, Oswald. 1982. Produktivität in der Wortbildung von Korpussprachen: Möglichkeiten der Heuristik. *Folia Linguistica* 16.225–39.

Perea, Manuel, Reem Abu Mallouh, and Manuel Carreiras. 2014. Are Root Letters Compulsory for Lexical Access in Semitic Languages? The Case of Masked Form-Priming in Arabic. *Cognition* 132.491–500.

Peters, Martin. 1980. *Untersuchungen zur Vertretung der indogermanischen Laryngale im Griechischen. Sitzungsberichte der Österreichischen Akademie der Wissenschaften, philosophische-historische Klasse*, vol. 337. Wien: Österreichischen Akademie der Wissenschaften.

————. 1999. Ein tiefes Problem. In Heiner Eichner, Hans Christian Luschützky and Velizar Sadovski (eds.), *Compositiones Indogermanicae in memoriam Jochem Schindler*, 447–56. Praha: Enigma Corporation.

Pierrehumbert, Janet, and Mary E. Beckman. 1988. *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Pinker, Steven, and Alan Prince. 1988. On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition. *Cognition* 28.73–193.

Plag, Ingo. 1999. *Morphological Productivity: Structural Constraints in English Derivation*. Berlin: Mouton de Gruyter.

————. 2007. Review of Baayen and Schreuder 2003. *Language* 83.196–9.

Plag, Ingo, Christiane Dalton-Puffer, and Harald Baayen. 1999. Morphological Productivity across Speech and Writing. *English Language and Linguistics* 3.209–228.

Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt, and Michael Becker. 2010. Harmonic Grammar with Linear Programming: From Linear Systems to Linguistic Typology. *Phonology* 27.1–41.

Powell, James Thomas. 1988. Homeric Hapax Legomena and Other Infrequent Words. Ph.D. diss., Yale University.

Pozdniakov, Konstantin, and Guillaume Segerer. 2007. Similarity Place Avoidance: A Statistical Universal. *Linguistic Typology* 11.307–48.

del Prado Martín, Fermín Moscoso, Raymond Bertram, Tuomo Häikiö, Robert Schreuder, and R. Harald Baayen. 2004. Morphological Family Size in a Morphologically Rich Language: The Case of Finnish Compared with Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30.1271–8.

del Prado Martín, Fermín Moscoso, Avital Deutsch, Ram Frost, Robert Schreuder, Nivja H. de Jong, and R. Harald Baayen. 2005. Changing Places: A Cross-Language Perspective on Frequency and Family Size in Dutch and Hebrew. *Journal of Memory and Language* 53.496–512.

Prasada, Sandeep, and Steven Pinker. 1993. Generalization of Regular and Irregular Morphological Patterns. *Language and Cognitive Processes* 8.1–56.

Press, William, Saul Teukolsky, William Vetterling, and Brian Flannery. 1992. *Numerical Recipes in C: The Art of Scientific Computing*, 2nd edn. Cambridge: Cambridge University Press.

Prince, Alan, and Paul Smolensky. 1993 [2002]. *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report TR-2, Cognitive Science Center, Rutgers University. URL `http://roa.rutgers.edu/files/537-0802/537-0802-PRINCE-0-0.PDF`.

Pring, Julian Talbot (ed.). 1982. *The Oxford Dictionary of Modern Greek: Greek-English and English-Greek*. Oxford: Clarendon Press.

Probert, Philomen. 2003. *A New Short Guide to the Accentuation of Ancient Greek*. London: Bristol Classical Press.

————. 2006a. Accentuation in Ancient Greek Deverbative *ā*-stems. *Oxford University Working Papers in Linguistics, Philology, and Phonetics* 11.122–42.

————. 2006b. *Ancient Greek Accentuation. Synchronic Patterns, Frequency Effects, and Prehistory*. Oxford: Oxford University Press.

————. 2010. Ancient Greek Accentuation in Generative Phonology and Optimality Theory. *Language and Linguistics Compass* 4.1–26.

Raffelsiefen, R. 1999. Phonological Constraints on English Word Formation. In G. E. Booij and Jaap van Marle (eds.), *Yearbook of Morphology 1998*, 225–87. Dordrecht: Kluwer.

Rakoczy, Hannes, Annette Clüver, Liane Saucke, Nicole Stoffregen, Alice Gräbener, Judith Migura, and Josep Call. 2014. Apes Are Intuitive Statisticians. *Cognition* 131.60–8.

Rau, Jeremy. 2009. *Indo-European Nominal Morphology: The Decads and the Caland System*. Innsbruck: Innsbrucker Beiträge zur Sprachwissenschaft.

Revithiadou, Anthi. 1999. Headmost Wins: Head Dominance and Ideal Prosodic Form in Lexical Accent Systems. Ph.D. diss., Leiden University.

Reynolds, L. D., and N. G. Wilson. 1991. *Scribes and Scholars. A Guide to the Transmission of Greek and Latin Literature*, 3rd edn. Oxford: Clarendon Press.

Risch, Ernst. 1974. *Wortbildung der Homerischen Sprache*, 2nd edn. Berlin: de Gruyter.

Rix, Helmut. 1976 [1992]. *Historische Grammatik des Griechischen*, 2nd edn. Darmstadt: Wissenschaftliche Buchgesellschaft.

Rix, Helmut, Martin Joachim Kümmel, Thomas Zehnder, Reiner Lipp, and Brigitte Schirmer. 2001. *Lexikon der indogermanischen Verben: Die Wurzeln und ihre Primärstammbildungen*, 2nd edn. Wiesbaden: Reichert.

Ross, Sheldon. 2010. *A First Course in Probability*, 8th edn. Upper Saddle River, NJ: Prentice Hall.

Rumelhart, David E., and Jay L. McClelland. 1986. On Learning the Past Tenses of English Verbs. In *Paralell Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 2, 216–71. Cambridge, MA: MIT Press.

Saffran, Jenny, Elissa L. Newport, and Richard N. Aslin. 1996. Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language* 35.606–21.

Saffran, Jenny R. 2002. Constraints on Statistical Language Learning. *Journal of Memory and Language* 47.172–96.

————. 2003. Statistical Language Learning: Mechanisms and Constraints. *Current Directions in Psychological Science* 12.110–4.

Säily, Tanja. 2008. Productivity of the Suffixes *-NESS* and *-ITY* in 17th-Century English Letters: A Sociolinguistic Approach. MA thesis, University of Helsinki.

Sandell, Ryan. 2011a. Reduplication and Grammaticalization in Vedic Sanskrit. URL `https://www.academia.edu/1693673/Reduplication_and_Grammaticalization_in_Vedic_Sanskrit`, Presentation. 20th International Conference of Historical Linguistics, Ōsaka, Japan, July 28.

————. 2011b. The Morphophonology of Reduplicated Presents in Vedic and Indo-European. In Stephanie Jamison, H. Craig Melchert and Brent Vine (eds.), *Proceedings of the 22nd UCLA Indo-European Conference*, 223–54. Bremen: Hempen.

———. 2013. Vedic Perfect Weak Stems of the Form $C_1eC_2$-. URL https://www.academia.edu/5572089/Vedic_Perfect_Weak_Stems_of_the_Form_C1eC2, Presentation. 25th Annual UCLA Indo-European Conference, October 26.

———. 2014a. Compensatory Lengthening in Vedic and the Outcomes of Proto-Indo-Iranian *[az] and *[až]. In Stephanie W. Jamison, H. Craig Melchert and Brent Vine (eds.), *Proceedings of the 25th UCLA Indo-European Conference*, 183–202. Bremen: Hempen.

———. 2014b. The Phonological Origins of Indo-European Long-Vowel ("Narten") Presents. URL https://www.academia.edu/8858713/The_Phonological_Origins_of_Indo-European_Long-Vowel_Narten_Presents, Presentation. 26th UCLA Indo-European Conference, October 24–25.

Sandell, Ryan, and Sam Zukoff. 2014. A New Approach to the Origin of Germanic Strong Preterites. URL https://www.academia.edu/9052823/Sandell_and_Zukoff_2014_-_A_New_Approach_to_the_Origin_of_Germanic_Strong_Preterites_poster_, Poster Presentation. 45th Annual Meeting of the North East Linguistic Society, November 1.

de Saussure, Ferdinand. 1983. *Course in General Linguistics*. London: G. Duckworth.

Sauzet, P. 1989. L'Accent du grec ancien et les relations entre structure métrique et représentation autosegmentale. *Langages* 24.81–113.

Scalise, Sergio. 1984. *Generative Morphology*. Dordrecht: Foris.

Schindler, Joachem. 1975. Zum Ablaut der neutralen *s*-Stämme des Indogermanischen. In Helmut Rix (ed.), *Flexion und Wortbildung. Akten der V. Fachtagung der Indogermanischen Gesellschaft. Regensburg, 9.-14. September 1973*, 259–67. Reichert: Wiesbaden.

Schreuder, Robert, and R. Harald Baayen. 1995. Modeling Morphological Processing. In L. B. Feldman (ed.), *Morphological Aspects of Language Processing*, 131–54. Hillsdale, NJ: Erlbaum.

———. 1997. How Complex Simplex Words Can Be. *Journal of Memory and Language* 37.118–39.

Schultink, Henk. 1961. Produktiviteit als morphologisch phenomeen. *Forum der Letteren* 2.110–25.

Schwyzer, Eduard. 1938 [1953]. *Griechische Grammatik. Erster Band. Allgemeiner Teil. Lautlehre. Wortbildung. Flexion*, 4th edn. München: C. H. Beck.

di Sciullo, Anna Maria, and Edwin Williams. 1987. *On the Definition of Word*. Cambridge, MA: MIT Press.

Shettleworth, Sara J. 1998. *Cognition, Evolution, Learning*. Oxford: Oxford University Press.

Skousen, Royal. 1989. *Analogical Modeling of Language*. Dordrecht: Kluwer.

———. 2002. Introduction. In Royal Skousen, Deryle Lonsdale and Dilworth B. Parkinson (eds.), *Analogical Modeling: An Exemplar-Based Approach to Language*, 1–8. Amsterdam: John Benjamins.

Skousen, Royal, Deryle Lonsdale, and Dilworth B. Parkinson (eds.). 2002. *Analogical Modeling: An Exemplar-Based Approach to Language*. Amsterdam: John Benjamins.

Smolensky, Paul, and Géraldine Legendre. 2006. *The Harmonic Mind*. Cambridge, MA: MIT Press.

Snell, Bruno (ed.). 1979-2010. *Lexikon des frühgriechischen Epos*. Göttingen: Vandenhoeck & Ruprecht.

Steriade, Donca. 1988a. Greek Accent: A Case for Preserving Structure. *Linguistic Inquiry* 19.271–314.

———. 1988b. Reduplication and Syllable Transfer in Sanskrit and Elsewhere. *Phonology* 5.73–155.

———. 2014. A Synchronic Analysis of Ancient Greek Accent. Presentation. Harvard GSAS Colloquium, 22 September 2014.

van Strien-Gerritsen, Magdalena. 1973. *De homerische composita*. Assen: van Gorcum & Compagnie.

Strunk, Klaus. 1967. *Nasalpräsentien und Aoriste. Ein Beitrag zur Morphologie des Verbums im Indo-Iranischen und Griechischen*. Heidelberg: Carl Winter.

Sturtevant, Edgar H. 1947. *An Introduction to Linguistic Science*. New Haven: Yale University Press.

Sukhtankar, V. S., S. K. Belvalkar, S. K. De, and R. N. Dandekar (eds.). 1971–1975. *The Mahābhārata: Text as Constituted in its Critical Edition*. Poona: Bhandarkar Oriental Research Institute.

Sutherland, Norman S., and Nicholas J. Mackintosh. 1971. *Mechanisms of Animal Discrimination Learning*. New York: Academic Press.

Szemerényi, Oswald. 1966. The Origin of Vedic 'imperatives' in -si. *Language* 42.1–7.

Taft, Marcus, and Kenneth I. Forster. 1975. Lexical Storage and Retrieval of Prefixed Words. *Journal of Verbal Learning and Behavior* 14.638–47.

Tebben, Joseph R. 1994. *Concordantia Homerica. Pars I. Odyssea. A Computer Concordance to the Van Thiel Edition of Homer's Odyssey*. Hildesheim: Georg Olms.

Tesar, Bruce, and Paul Smolensky. 2000. *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.

van Thiel, Helmut (ed.). 1991. *Homeri Odyssea*. Hildesheim: Georg Olms.

————. 1996. *Homeri Ilias*. Hildesheim: Georg Olms.

Thompson, L. C., and M. T. Thompson. 1996. *Thompson River Salish Dictionary*. No. 16 in University of Montana Occasional Papers in Linguistics. Missoula, Montana: University of Montana, Missoula.

Topintzi, Nina. 2010. *Onsets*. Cambridge: Cambridge University Press.

Traficante, Daniela, and Cristina Burani. 2003. Visual Processing of Italian Verbs and Adjectives. In R. Harald Baayen and Robert Schreuder (eds.), *Morphological Structure in Language Processing*, 45–64. Berlin: Mouton de Gruyter.

Tsujimura, Natsuko. 2014. *An Introduction to Japanese Linguistics*. Malden, MA: Wiley-Blackwell.

Tucker, Elizabeth Fawcett. 1988. Some Innovations in the System of Denominative Verbs in Early Indic. *Transactions of the Philological Society* 86.93–114.

————. 1990. *The Creation of Morphological Regularity: Early Greek Verbs in -éō, -áō, -óō, -úō, and -íō*. Göttingen: Vandenhoeck & Ruprecht.

Underwood, B. J. 1983. *Attributes of Memory*. Glenview, IL: Scott, Foresman.

van Nooten, Barend, and Gary Holland (eds.). 1995. *Rig Veda. A Metrically Restored Text*. Cambridge, MA: Harvard University Press.

Vance, Timothy J. 1987. *An Introduction to Japanese Phonology*. Albany: State University of New Yorke.

Vannest, Jennifer, Elissa L. Newport, Aaron J. Newman, and Daphne Bavelier. 2011. Interplay between Morphology and Frequency in Lexical Access: The Case of the Base Frequency Effect. *Brain Research* 1373.144–59.

Vendryes, Joseph. 1904 [1945]. *Traité d'accentuation grecque*. Paris: Klincksieck.

Vine, Brent. 2007. Latin *gemō* 'groan', Greek γέγωνε 'cry out', and Tocharian A *ken-* 'call'. In Alan Nussbaum (ed.), *Verba Docenti. Studies in Historical and Indo-European Linguistics Presented to Jay H. Jasanoff by Students, Colleagues, and Friends*, 343–58. Ann Arbor: Beech Stave.

Wackernagel, Jacob. 1896. *Altindische Grammatik I. Lautlehre*. Göttingen: Vandenhoeck and Ruprecht.

———. 1905. *Altindische Grammatik II. Einleitung zur Wortlehre. Nominalkomposition*. vol. II. Göttingen: Vandenhoeck & Ruprecht.

Wackernagel, Jacob, and Albert Debrunner. 1954. *Altindische Grammatik. II.2 Die Nominalsuffixe*. Göttingen: Vandenhoeck & Ruprecht.

Watkins, Calvert. 1962. *Indo-European Origins of the Celtic Verb I. The Sigmatic Aorist*. Dublin: Dublin University Press.

West, Martin. 1982. *Greek Metre*. Oxford: Clarendon Press.

West, Martin L. (ed.). 1998. *Homeri Ilias*. Stuttgart & Leipzig: Teubner.

Westcott, Brooke Foss, and Fenton Anthony Hort (eds.). 1881. *The New Testament in the Original Greek*. Cambridge: Macmillan.

Whitney, William Dwight. 1885 [1963]. *The Roots, Verb-Forms, and Primary Derivatives of the Sanskrit Language*. Delhi: Motilal Banarsidass.

Widmer, Paul. 2004. *Das Korn des weiten Feldes. Interne Derivation, Derivationskette, und Flexionsklassenhierarchie: Aspekte der nominalen Wortbildung im Urindogermanischen*. Innsbruck: Innsbrucker Beiträge zur Sprachwissenschaft.

Witzel, Michael. 2003. *Das alte Indien*. München: C. H. Beck'sche Verlagsbuchhandlung.

Wodtko, Dagmar S., Britta Irslinger, and Carolin Schneider. 2008. *Nomina im Indogermanischen Lexicon*. Heidelberg: Carl Winter.

WordHoard. 2004–2011. WordHoard. An Application for the Close Reading and Scholarly Analysis of Deeply Tagged Texts. URL `www.wordhoard.northwestern.edu`.

Wurm, Lee H. 1997. Auditory Processing of Prefixed English Words is Both Continuous and Decompositional. *Journal of Memory and Language* 37.438–61.

Zacks, Rose T., and Lynn Hasher. 2002. Frequency Processes: A Twenty-Five Year Perspective. In Peter Sedlmeier and Tilmann Betsch (eds.), *Frequency Processing and Cognition*, 21–36. Oxford: Oxford University Press.

Zacks, Rose T., Lynn Hasher, and Henrianne Sanft. 1982. Automatic Encoding of Event Frequency: Further Findings. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 8.106–16.

Zeldes, Amir. 2012. *Productivity in Argument Selection. From Morphology to Syntax*. Berlin: de Gruyter.

Zipf, George Kingsley. 1935. *The Psycho-Biology of Language*. Boston: Houghton Mifflin.

Zukoff, Sam. 2014. On the Origins of Attic Reduplication. In Stephanie W. Jamison, H. Craig Melchert and Brent Vine (eds.), *Proceedings of the 25th UCLA Indo-European Conference*, 257–78. Bremen: Hempen.

———. 2015. Poorly-Cued Repetition Avoidance in Indo-European Reduplication. Presentation. 89th Annual Meeting of the Linguistic Society of America, January 9.

Zuraw, Kie. 2000. Patterned Exceptions in Phonology. Ph.D. diss., University of California, Los Angeles.

———. 2009. Frequency Influences on Rule Application within and across Words. In *Proceedings of the Chicago Linguistic Society*, vol. 43, 283–309. Chicago: University of Chicago Press.

Zwicky, Arnold. 1985. Heads. *Journal of Linguistics* 21.1–29.

———. 1993. Heads, Bases, and Functors. In Greville G. Corbett, N. Fraser and S. McGlashan (eds.), *Heads in Grammatical Theory*, 292–315. Cambridge: Cambridge University Press.