

The Identification of Bases in Morphological Paradigms

Adam C. Albright

University of California, Los Angeles

June, 2002

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Linguistics

UCLA Linguistics Department

Thesis committee:

Bruce Hayes, co-chair

Donca Steriade, co-chair

Carson Schütze

Brent Vine

© Copyright by
Adam C. Albright
2002

Contents

1	Introduction	1
1.1	The problem of bases in word-based morphology	8
1.2	The problem of bases in stem-based morphology	13
1.3	The problem of bases in correspondence theory	14
1.4	Plan of the thesis	15
2	Paradigm leveling in Yiddish	17
2.1	Leveling to the 1sg in Yiddish verb paradigms	18
2.1.1	MHG present tense patterns	18
2.1.2	Yiddish present tense patterns	19
2.2	The 1sg as the “optimal base” in Yiddish	23
2.2.1	Identifying the optimal base	23
2.2.2	Using the 1sg as the base of Yiddish verb paradigms	28
2.3	Comparison with other German dialects	30
2.4	Local summary	31
2.5	Significance of the single surface base hypothesis	32
3	Identifying bases algorithmically	35
3.1	Desiderata for an automated learner	36
3.2	The minimal generalization learner	38
3.3	Metrics for sub-grammars	43
3.4	Results for synthetic languages	45
3.4.1	Synthetic language 1: Neutralization in suffixed forms	45
3.4.2	Synthetic language 2: Neutralization in suffixless forms	47
3.4.3	Synthetic language 3: Morphological neutralization	48
3.4.4	Synthetic language 4: Lexical exceptions	50
4	The Latin <i>honor</i> analogy	53
4.1	Statement of the problem	54
4.2	Applying the model to Latin noun paradigms	59
4.3	Projecting nominatives from the genitive	65
4.4	Discussion	68
4.4.1	Frequency of occurrence within the paradigm	69
4.4.2	Token frequency of different paradigm members	70
4.4.3	Semantics and local markedness	71

4.4.4	Leveling vs. extending alternations	72
4.4.5	Local summary	73
5	Extension of <i>ablaut</i> in Lakhota	75
5.1	Background on Lakhota	78
5.1.1	Phoneme inventory and phonotactics	78
5.1.2	Final vowel alternations (“ablaut”)	78
5.1.3	Reduplication	81
5.2	Innovative paradigms in Lakhota	83
5.3	Restricting UR discovery to a single surface form	85
5.3.1	What is the most informative part of Lakhota paradigms?	86
5.3.2	Consequences of using a 2nd person form as the UR	92
5.4	Inferring nothing from non-basic forms	94
5.5	Other examples of inconsistent paradigms	95
5.6	Local summary	96
6	Discussion	99
6.1	URs	100
6.2	Bases of analogical change	106
6.2.1	Why a nominative/3sg/etc. preference?	107
6.2.2	Local Markedness	114
6.2.3	Why a leveling preference?	115
6.2.4	Frequency, forgetting, and overregularization	116
6.3	Scaling up: discovering local bases	118
6.3.1	The need for local bases	118
6.3.2	Extending the model to find local bases	121
6.4	Relation to acquisition evidence	126
6.5	Summary	127
7	Conclusion	129
Appendix A	Results for the Synthetic <i>Neutralize I</i> Language	133
A.1	Input forms	133
A.2	Absolute to ergative direction	133
A.2.1	Grammar for the <i>absolute</i> → <i>ergative</i> direction	133
A.2.2	Using the grammar to derive ergatives for each word	142
A.2.3	Calculation of metrics in the <i>absolute</i> → <i>ergative</i> direction	142
A.3	Ergative to absolute direction	143
A.3.1	Grammar for the <i>ergative</i> → <i>absolute</i> direction	143
A.3.2	Using the grammar to derive absolutes for each word	152
A.3.3	Calculation of metrics in the <i>ergative</i> → <i>absolute</i> direction	153
Appendix B	Metrics for base selection in Latin	155
References		157

List of Figures

1.1	Routes for deriving different types of words	12
2.1	Summary of neutralizations affecting German verb forms	31
2.2	Summary of neutralizations affecting Yiddish verb forms	31
3.1	Comparison of different candidates for base status	35
3.2	Time course of acquisition of the minimal generalization model	43
4.1	Comparison of potential bases according to the criteria in §3.3	64
4.2	Average rank of forms as potential bases	64
4.3	Preference for <i>-r</i> or <i>-s</i> in different categories of nouns	66
4.4	Comparison of agentive vs. non-agentive nouns	67
5.1	An asymmetrical neutralization	75
5.2	A symmetrical neutralization	75
5.3	Innovative forms are incompatible with any UR	85
5.4	Comparing possible sources of Lakhota URs	91
6.1	Sample input assuming unequal frequencies	109
6.2	Theoretical neutralization of word-final obstruents in Korean	110
6.3	Actual neutralization of word-final obstruents in Korean	111
6.4	Routes for deriving different types of words	117
6.5	Three ways to produce a non-basic form using a local base	120
6.6	Setting up local bases by finding pockets of high reliability	121
6.7	Result of “local demotion” for Latin noun paradigms	122
6.8	Results of global base selection for Spanish present tenses	123
6.9	Circularity and ties in the Spanish present tense paradigm	123
6.10	3sg as local base of Spanish paradigm	124
6.11	3pl, 3sg as local bases of Spanish paradigm	124

List of Tables

1.1	Comparison of factors encouraging the extension of a form	7
2.1	Generalized form is not the 2,3sg	20
2.2	Summary of neutralizations in Yiddish	28
4.1	Distribution of singular tokens for some Latin nouns	71
5.1	Lakhota consonant inventory	79
5.2	Lakhota vowel inventory	79
5.3	Number of words in each ablaut category	87
5.4	Lakhota subject markers	89
5.5	Number of words in each agreement pattern	90
6.1	Distribution of terminations and accent patterns	104
6.2	Percent correct in deriving Latin noun paradigms	122
6.3	Percent correct in deriving Spanish present tense paradigms	123
6.4	Competition for basehood of 1sg,2sg,3sg,3pl subparadigm	125
B.1	Base selection metrics for Latin noun paradigms	155

ABSTRACT OF THE DISSERTATION

The Identification of Bases in Morphological Paradigms

by

Adam C. Albright
Doctor of Philosophy in Linguistics
University of California, Los Angeles, 2002
Professor Bruce Hayes, Co-chair
Professor Donca Steriade, Co-chair

Many theories, in many domains of linguistics, assume that some members of morphological paradigms are more basic than others. Bases of paradigms are privileged in various ways: they may determine phonological properties of other forms, they may determine the direction of analogical changes, and so on. In this thesis, I propose that such effects are a result of the procedure by which learners seek to develop a grammar that allows them to project inflected forms as accurately and confidently as possible. I present a computationally implemented model of paradigm acquisition that attempts to use one form in the paradigm as the base to project the remaining forms, using stochastic morphological rules. I pursue two hypotheses about how this is done. The first is that learners are limited to selecting a single form as the base, and that the base must be a surface form from somewhere within the paradigm. Furthermore, the choice of base is global, meaning that the same slot must serve as the base for all lexical items. The second hypothesis is that learners select the base form that is maximally informative, in the sense that it preserves the most contrasts, and permits accurate productive generation of as many forms of as many words as possible.

As evidence for this approach, I analyze three cases in which an typologically marked form served as the base of a historical analogical change: Yiddish present tense paradigms (in which all forms were remodeled on the 1st sg), Latin noun paradigms (in which nominatives were remodeled on oblique forms), and Lakhota verbs (in which unsuffixed forms are being remodeled on suffixed forms). In each case, I show how the model correctly selects the base form, and also correctly predicts asymmetries in the direction of subsequent paradigmatic changes. I show that these asymmetries are not predicted by a more traditional model of underlying forms, in which learners compare all of the parts of the paradigm to construct abstract underlying representations that combine unpredictable information from multiple forms. Finally, I discuss possible extensions of this model to accommodate larger paradigms with multiple, local bases.

Acknowledgments

I would like to thank my committee members (Bruce Hayes, Donca Steriade, Carson Schütze and Brent Vine) for all of their patience, encouragement, and friendly criticism, and especially for agreeing to show up for a defense at 9am on a Sunday morning. First and foremost, I am thankful to Bruce Hayes, who has been a patient and wise teacher, an attentive advisor, and an enthusiastic collaborator during my time at UCLA. He has helped me more in my research and growth than many pages of acknowledgments could do justice to. I am also greatly indebted to Donca Steriade, for her wisdom and guidance, and for her knack for making hard problems seem so clear, at least until the time when I left her office. I am equally grateful to Carson Schütze for our many stimulating and challenging discussions in which he pushed me to take on hard issues, clarify my analysis, and see things from another angle. Finally, I am much obliged to Brent Vine for his thoughtful and insightful questions, for offering intriguing parallels and additional data, and for clearing up at least a few of my many misconceptions about Latin and Russian.

I would also like to thank all of the other members of the UCLA Linguistics Department who have made it such a vibrant and educational environment. I owe particular thanks to Colin Wilson for his discussions and advice, and to Pamela Munro for introducing me to Lakhota, and for generously providing me with her insights and the use of her Lakhota verb list. I am also especially grateful to Victoria Anderson, Marco Baroni, Rebecca Brown, Leston Buell, Ivano Caponi-gro, Taehong Cho, Christina Foreman, John Foreman, Matt Gordon, Amanda Jones, Sun-Ah Jun, Pat Keating, Sahyang Kim, Peter Ladefoged, Harold Torrence, Motoko Ueyama, Jie Zhang, Kie Zuraw, and all of the other members of the UCLA Phonetics Lab and Linguistics department (too numerous to name) who have made my years here stimulating and entertaining. In addition, I would like to thank my teachers during my undergraduate years at Cornell who instilled me with a love of phonology and historical linguistics—especially Abby Cohn, Wayne Harbert, and Jay Jasanoff.

Much of the data presented here would have been difficult or impossible for me to collect or make sense of on my own; I am particularly indebted to my Lakhota teacher Mary Rose Iron Teeth, for many hours patiently and cheerfully teaching me about her language. I am also extremely grateful to Angelo Mercado for his generous Latin help, and likewise to Jongho Jun, Taehong Cho, and Sahyang Kim for Korean, and to Katya Pertsova for Russian.

Various portions of this dissertation have benefited from questions and comments by audiences at UCLA, MIT, BU, the LSA (Chicago, January 2001), the ZAS Conference on Paradigm Uniformity (Berlin, March 2001), BLS 28 (Berkeley, February 2002), WCCFL 21 (Santa Cruz, April 2002), and the GLOW Phonological Acquisition Workshop (Utrecht, April 2002).

This dissertation work was supported in part by an NSF Graduate Fellowship, NSF grant

BCS-9910686, and a UCLA Dissertation Year Fellowship, all of which I gratefully acknowledge.

Writing a dissertation can be an ordeal, and it is made tolerable only by the unswerving support and encouragement of friends—even when you appear to have momentarily abandoned them. For this, I am eternally grateful to Moto and Marco, to Jie, Ivano, and Harold, to Scott, and most of all to Angelo. Last but not least, I am thankful to my parents, for their years of support and for giving me my first introduction to language.