THE UNIVERSITY of EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

PhD Centre for Doctoral Training in Data Science

Rules, Frequency, and Predictability in Morphological Generalization: Behavioral and Computational Evidence from the German Plural System

by

Katherine (Kate) McCurdy

2024

Declaration

This project report is submitted in partial fulfilment of the requirements for the degree of PhD Centre for Doctoral Training in Data Science. I declare that this thesis was composed by myself, that the work contained therein is my own, except where explicitly stated otherwise in the text, and that it has not been submitted, in whole or in part, for any other degree or professional qualification.

Katherine (Kate) McCurdy

Word Count: 55107 words

This thesis was conducted under the supervision of Dr. Adam Lopez and Dr. Sharon Goldwater.

Abstract

Morphological generalization, or the task of mapping an unknown word (such as a novel noun *Raun*) to an inflected form (such as the plural *Rauns*), has historically proven a contested topic within computational linguistics and cognitive science, e.g. within the past tense debate (Rumelhart and McClelland, 1986; Pinker and Prince, 1988; Seidenberg and Plaut, 2014). Marcus et al. (1995) identified German plural inflection as a key challenge domain to evaluate two competing accounts of morphological generalization: a *rule generation* view focused on linguistic features of input words, and a *type frequency* view focused on the distribution of output inflected forms, thought to reflect more domain-general cognitive processes. More recent behavioral and computational research developments support a new view based on *predictability*, which integrates both input and output distributions. My research uses these methodological innovations to revisit a core dispute of the past tense debate: how do German speakers generalize plural inflection, and can computational learners generalize similarly?

This dissertation evaluates the rule generation, type frequency, and predictability accounts of morphological generalization in a series of behavioral and computational experiments with the stimuli developed by Marcus et al.. I assess predictions for three aspects of German plural generalization: distribution of infrequent plural classes, influence of grammatical gender, and within-item variability. Overall, I find that speaker behavior is best characterized as frequency-matching to a phonologically-conditioned lexical distribution. This result does not support the rule generation view, and qualifies the predictability view: speakers use some, but not all available information to reduce uncertainty in morphological generalization. Neural and symbolic model predictions are typically overconfident relative to speakers; simple Bayesian models show somewhat higher speaker-like variability and accuracy. All computational models are outperformed by a static phonologically-conditioned lexical baseline, suggesting these models have not learned the selective feature preferences that inform speaker generalization.

Lay Summary

This dissertation investigates how human speakers handle novelty in language. When a speaker hears a new word, they sometimes need to guess which linguistic category it belongs to; for example, if an English speaker wants to use the plural form of a new word “Raun,” they need to decide whether its plural form is more likely to be “Rauns” (like “dogs”), or “Raun” (like “sheep”). This guessing task is even more complicated in German, where the plural could be “Rauns,” “Raun,” “Raunen,” “Raune,” or “Räuner.”

Linguists have developed several theories as to the key factors that influence how German speakers create new plural words. Perhaps the main factor is the statistical frequency of a plural ending, or how many different kinds of words it combines with, or some other important property of the word such as whether it is masculine, feminine, or neuter. Researchers have also conducted behavioral experiments directly asking speakers to pluralize words like “Raun,” or used computational models trained to predict plurals. Still, findings have been inconclusive.

In this thesis, I test how German speakers produce the plural form of invented test words like “Raun” in a range of experiments. I then compare their behavior to the predictions of various computational models, including artificial neural network models trained with deep learning, and statistical models which learn to apply symbolic rules such as “if the noun has feminine gender, then predict the *-en* plural class.”

Overall, I find that German speakers are most influenced by the frequency of the plural ending, and only somewhat influenced by the properties of individual words: they consistently produce a similar range of plural forms for each test word. Most computational models, however, tend to confidently predict a specific plural ending based on the individual properties of a test word, such as whether it has masculine or feminine gender. This discrepancy suggests that, if we want computational models to handle linguistic novelty in human-like ways, in some cases we’ll want frequency-matched predictions which disregard apparently informative cues.

Dedication

With gratitude to the many whose support and compassion I've benefited from along the way, and preemptive apologies for a necessarily partial list here. To my dissertation supervisors, Adam Lopez and Sharon Goldwater, whose relentless dedication to writing and thinking clearly has strengthened my research by some order of magnitude. To my dissertation examiners, Chris Dyer and Frank Keller, for close reading and a thoughtful and challenging viva discussion. To my internship supervisor, Paul Smolensky, for kindly welcoming my small contribution to a brilliant research career, and not gloating too much about beating the summer interns at beach volleyball. To my earlier academic supervisors and mentors, in reverse chronological order: Shravan Vasishth, Jesse Snedeker, Michael Wagner, Heather Goad — thank you all for training me how to do research, giving me the most stellar preparation to pursue a PhD, watching me disappear to work at a startup, then dusting off your reference letters for me years later when I finally heeded the siren call of academia. On that note, special thanks to two dear friends who were instrumental in that process (part of the siren chorus?): a 2017 road trip visit to Jean Yang accidentally tumbled me on to the doctoral candidate interview circuit, and then Tim O'Donnell's invaluable advice saw me through the mounting pressure of grad school applications. All of you made me a researcher, and I am grateful for it.

Academia's siren call would have wrecked me, were it not for the brilliant and kind shipmates I found among the graduate student community in Edinburgh — especially my irrepressible officemates and fellow mermaids Seraphina Goldfarb-Tarrant and Ida Szubert, who kept me tethered to sanity, joy, and wild mushrooms through the pandemic. But there were so many beacons in the storm: Sunday donut walks with Alex Robertson and Tanja Hendriks, radical utopian sci-fi with Naomi Saphra, pre-pandemic Montréal jazz dancing with Janie Sinclair, post-pandemic northern soul dancing with Drew Altschul and Ella Edginton, singing the revolution with Cailean Gallagher and the communist choir, underground comedy with Julie-Anne

Meaney, linguistic gossip dinners with Henry Conklin, bougie salad lunches with Lucy Havens, and Tom Sherborne's beautiful lockdown Christmas vegan Wellington. Within the university, I've had the fortune of multiple wonderful academic homes. My thanks to the Agora research group, especially Yevgen Matusevych, Elizabeth Nielsen, Ramon Sanabria, Andreas Grivas, and Coleman Haley; to Neil Bramley and the CocoSci research group; to Itamar Kastner and the EdinMorph seminar attendees; to Kenny Smith and the CLE, whose collective work profoundly influenced this dissertation; to the incredible ILCC and Informatics faculty as a whole, particularly Chris Lucas, Nadin Kokciyan, Frank Mollica, and Siddharth Narayanaswamy for candid after-hours chat; and to the CDT Data Science program, including my fantastic 2018 cohort-mates — Markus (and Nicole, and Emilia!), Sigrid, Kasia, Will, Kaan, Aidan, Linus, Eleanor, Julie-Anne, Tom, and Ondrej — and our cohort elders, especially James Owers, Maria Gorinova, Clara Vania, Matt Rounds, James Garforth, and Ben Rhodes. Thanks as well to the institutional, administrative, and janitorial staff who make this learning environment possible, especially our nonpareil original program secretary Sally Galloway; Alex Lascarides and Amos Storkey for moving heaven and earth, bureaucratically speaking, so that I could run the first ever ILCC/CDT Writing Retreat in spring 2022; and John the Cleaner for bits of banter to lighten late-working evenings. All of you made the university into a wonderful community, and I am grateful to have been a part of it.

Throughout my doctoral studies, I've had fantastic collaborators on all sorts of projects. My thanks to Karen Gregory for building interdisciplinary bridges with the Digital Worker Inquiry event; to the Neurocompositional AI group at Microsoft Research and Johns Hopkins, especially Paul Soulos; and to all the co-authors of our award-wining 2021 CoNLL paper, especially our formidable lead Verna Dankers. Thanks as well to my current collaborator and post-doc supervisor Michael Hahn, who has shown exemplary patience as I've finished up my thesis corrections. It has truly been a pleasure to work with all of you, and I am grateful to have had the opportunity.

Finally, in the times I had to turn back from the sirens, I found necessary solace in ports beyond academia. Thanks to my peerless crew of fearless once-and-forever Berliners, especially Julia Schubert, Ali Hugill and Dan Dorocic, Mana Mathana and Jillian York, Nico Leaver and Sohrob Kazerounian, and Andrew Bateman and Ashish Mahajan for keeping your doors (docks?) open to me, and for the singular company of Suvi Lensu, Evanna Folkenfolk, and

Charley Wu. I'm thankful as well for old Stateside friends who have kept close despite my itinerant ways, especially Andrew Lyubarsky, Isaac Ericson, Caroline Zimmerman, Angel Chen, Meredith Tamminga, and my oldest and dearest friend, Rachel Cannon. None of this journey would have been possible without my parents, Robert and Virginia McCurdy, who love and support their children in all their far-flung pursuits; my heartfelt thanks to them, my siblings Rob, Ryan, and Annie, and their partners Jasmine, Hyung-Ju, and Daniel, for warmth and grounding when I needed it. And thanks to Peter Frase for caring for me in so many ways, and being easy to be around. All of you make me who I am, and I'm grateful for it.

Contents

Declaration i

Abstract iii

Lay Summary v

Dedication ix

List of Figures xv

List of Tables xix

1 Introduction 1

- 1.1 Regularity and German plural inflection . . . 2
- 1.2 Shifting toward distributions . . . 3
- 1.3 Summary and research questions . . . 5
- 1.4 Present work . . . 6
- 1.5 Chapter Outline . . . 8

2 Background: Morphological Generalization 11

- 2.1 Regularity: Generalization and the lexicon . . . 12
  - 2.1.1 Defining morphological regularity . . . 12
  - 2.1.2 Information theory and the Low Conditional Entropy Conjecture . . . 17
  - 2.1.3 Defaults in linguistic theory . . . 20
  - 2.1.4 Type frequency . . . 24
- 2.2 Regularization: Generalization and speakers . . . 26
  - 2.2.1 Regularity and speaker generalization: Wug tests . . . 27

2.2.2 Regularization and speaker generalization in artificial languages 31
2.3 Computational models of morphological generalization 39
2.3.1 Symbolic models 40
2.3.2 Frequency-matching models 42
2.3.3 Exemplar-based and neural models 45
2.4 Chapter summary 50

3 Background: German Plural 53

3.1 The German noun lexicon 54
3.1.1 Lexical distributions 55
3.2 Linguistic theories of German plural inflection 59
3.2.1 Minority default 60
3.2.2 Gender-conditioned defaults 64
3.2.3 Schema theory 66
3.3 Behavioral evidence 67
3.3.1 Four ways to generalize 68
3.3.2 Köpcke (1988) and Marcus et al. (1995) 70
3.3.3 Other behavioral studies 77
3.4 Computational evidence 79
3.5 Conclusion: Research Questions 83

4 Speakers default to variation 85

4.1 Introduction 85
4.2 Study 1 87
4.2.1 Introduction 87
4.2.2 Study 1a: Speaker plural inflection 89
4.2.3 Study 1b: Encoder-Decoder inflection 93
4.2.4 Discussion 97
4.2.5 Conclusions 99
4.3 Additional Results 100
4.3.1 Modeling Experiments and Results 100
4.3.2 Speaker Data Analysis 106

4.4 Discussion . . . 110
4.4.1 No evidence for minority default in speaker plural generalization. 110
4.4.2 Models are overconfident relative to speakers. . . . 113
4.5 Conclusion . . . 118

5 Speakers don't really use grammatical gender 119

5.1 Introduction . . . 119
5.2 Study 2 . . . 120
5.2.1 Introduction . . . 120
5.2.2 Background . . . 121
5.2.3 Method . . . 122
5.2.4 Results . . . 123
5.2.5 Discussion . . . 127
5.3 Study 3 . . . 129
5.3.1 Behavioral experiment . . . 130
5.3.2 Results . . . 130
5.4 Discussion . . . 132
5.5 Conclusion . . . 134

6 Speakers probability-match and condition gender on phonology 135

6.1 Introduction . . . 135
6.2 Study 4 . . . 136
6.2.1 Motivation . . . 136
6.2.2 Background . . . 138
6.2.3 Methods . . . 140
6.2.4 Results and Discussion . . . 143
6.2.5 General Discussion . . . 146
6.2.6 Conclusion . . . 148
6.3 Discussion . . . 149
6.4 Conclusion . . . 151

7 Overview and Synthesis of Results 153
7.1 Additional Models of the German Noun Lexicon . . . 154
7.1.1 Bayesian Classifier and Naive Bayes . . . 154
7.1.2 Nearest Neighbors . . . 156
7.1.3 Additional Lexical Baselines . . . 157
7.2 Overall Model Comparison . . . 158
7.3 Summary of Findings . . . 162
7.4 Discussion . . . 165
7.5 Conclusion . . . 168

8 Conclusion 169
8.1 Contributions . . . 170
8.2 Limitations and Future Directions . . . 171
8.2.1 Behavioral . . . 171
8.2.2 Computational: Toward modeling speaker-like generalization . . . 173
8.3 Finis . . . 177

References 179

List of Figures

3.1 Distribution of plural classes for 25,999 German nouns in CELEX2 (Baayen et al., 1995), grouped by grammatical gender and the final character of the singular noun. Bubble size indicates number of nouns. Though the *-s* plural class is rare overall, it appears with 53 unique character-gender combinations, a more diverse set of environments than any other plural class. 57

3.2 Distribution shown in Fig. 3.1, factorized into the joint distribution of gender and final character ($p(x)$, above) and plural class conditional on those contexts ($p(y|x)$, below). Higher observed counts for input feature combinations (upper plot) are represented as higher opacity (lower plot). While some feature combinations are uniformly associated with a particular plural class (e.g. feminine nouns ending in *-e)*, others show more variable distributions (e.g. nouns ending in *-t*, or nonfeminine nouns ending in *-n*). 58

3.3 Average item-level ratings reported by Marcus et al. (1995). 75

3.4 The Plural-Computer, which assigns German nouns a regular plural form based on their gender (Genus) and final character/s (Endbuchstabe/n) (Mugdan, 1977, 53). 79

4.1 Plural class productions by item. 96

4.2 Distribution of plural classes by rank in ED model output. 96

4.3 Distribution of predicted plural class for Marcus et al. stimuli. Both the neural RNN Encoder-Decoder (ED) model and the symbolic ATP model predict *-s* more frequently for phonologically atypical nouns (“Non-Rhymes”) compared to typical stimuli (“Rhymes”). 104

4.4 Distribution of predicted plural class for Marcus et al. stimuli by item. Both the neural RNN Encoder-Decoder (ED) model and the symbolic ATP model predict *-s* almost exclusively for nouns ending in *-k*, which all happen to be phonologically atypical nouns. 105

4.5 Speaker productions from McCurdy, Goldwater and Lopez (2020), with phonologically atypical nouns further subdivided by ending with *-k*. Speakers produce *-s* more frequently for the 7 atypical nouns ending in *-k*; for the other 5 atypical nouns, speakers produce *-s* at rates equivalent to the 12 typical nouns. 107

4.6 Distribution of average item ratings by phonological typicality and presentation category (data from Marcus et al., 1995, Appx. 3). For *-s* plural forms (top row) in the Root and Borrowing categories (upper left / upper middle), average ratings for atypical nouns ending in *-k* are significantly higher than both typical nouns and atypical nouns with other endings. This *-k*-ending asymmetry does not appear for other plural endings (lower row). There is also no asymmetry for *-s* plural forms in the Name category (upper right), where ratings are generally high because *-s* is the required plural class for German proper names. 108

4.7 Percentage of productions per item for most frequent class for that item. "-" indicates a model trained on wordform alone, while "+ gender" indicates a model trained on wordform and grammatical gender. For any stimulus item, roughly 80-90% of the neural RNN ED and symbolic ATP model instances agree on the plural class for that item, while on average 45-50% of speakers assign the same class to one item. 116

5.1 Distribution of plural suffix overall (upper) and by gender (lower) in the UniMorph corpus. 121

5.2 Plural suffix productions by gender, speakers (lower) vs. ED model (upper) . . 124

5.3 Individual speaker variation in plural suffix production by gender. Each speaker saw 8 words from each gender, shown on the y-axis. For each gender and plural suffix, the boxes indicate the median and interquartile range of individual speaker productions for that combination. For all gender categories, the median number of *-e* productions is 4, while the median number of *-en* productions is 3. 125

5.4 Distribution of plural suffix by gender for consonant-final monosyllabic words in UniMorph . . . 128

5.5 Distribution of mutual information (MI) between plural class and grammatical gender per participant per study. Note that the slight increase in MI for S3 appears largely driven by three individual participants. . . . 131

6.1 Reference distributions calculated from German CELEX2 (Baayen et al., 1995). Top, all nouns (excluding ambiguous gender nouns). Bottom, monosyllabic consonant-final nouns. . . . 138

6.2 Task presentation for one item. To the left of the novel noun is gender selection, to the right, written plural. . . . 141

6.3 Gender and plural productions from participants. Compare to reference distributions in Fig. 6.1. . . . 143

6.4 Gender and plural productions by item. . . . 144

6.5 90% CIs for reference distributions (all and monoCF nouns), and observed values for speakers. Speakers may regularize overall variation in $H(C)$, but do not appear to regularize $MI(C;G)$. . . . 144

6.6 H(C) and MI(C;G) by participant. Color boxes and lines show 90% CI for all (red) and monoCF (green) nouns; color dots show reference values. The blue dot shows the speaker grand mean, and the blue line shows a Loess regression fit of speaker MI(C;G) on H(C). Most participants are in the green box, consistent with probability-matching the monoCF noun distribution. . . . 145

6.7 Distribution of mutual information (MI) between plural class and grammatical gender per participant per study. Observed values are similar across studies, with participants in S4 tending toward slightly *lower* gender conditioning on average. . . . 149

7.1 Overview: Jensen-Shannon divergence between speaker productions and model predictions for each stimulus item in each behavioral experiment. See model descriptions in Table 7.3. Boxplots summarize median values and interquartile range from items in all experiments, while the colored dots give separate mean values and standard error bars for each of the four behavioral experiments (Table 7.2). For ease of visualization, divergences greater than 0.3 bits are not plotted. A phonologically conditioned baseline (M95 Rhymes CLX; Table 7.1) shows the lowest median JSD, followed by the Naive Bayes classifier trained without gender. 161

7.2 Entropy per stimulus item and grammatical gender. Measurements for model predictions are plotted in points and smoothed lines, for models trained without gender (top panel) and with gender; measurements for speaker productions are shown in grey crossbars (max, median, min), including all gender combinations in top panel. Stimuli are ordered from highest (*Raun*) to lowest (*Fnöhk*) likelihood (c.f. Table 3.6). Trend lines are generated by loess (local nonparametric) regression. Note that only Bayesian model predictions (blue lines) are consistently within the entropy range of speaker productions (grey boxes). . . 164

List of Tables

2.1 Five criteria for regularity identified by Herce (2019), organized by two conceptual dimensions: 1) whether regularity is defined with respect to properties of linguistic categories considered in *isolation* as opposed to some lexical *distribution* over multiple categories, and 2) whether regularity is defined with respect to inputs, outputs, or their joint interaction. Productivity is listed separately, as it is the dependent variable for our purposes. 13

2.2 Paradigms illustrating three inflectional classes (C1–C3) in Swedish nouns, reproduced from Round et al. (2022). Morphological exponents are indicated with boldface. 17

2.3 Arrangement of Swedish nominal exponents from Table 2.2, according to inflection class (row) and grammatical feature (column). 17

2.4 Suffixes for German weak adjective inflection. German adjectives are inflected to agree with nouns in case, grammatical gender, and number. 21

2.5 Generative analysis of German weak adjective inflection (Zwicky, 1985). The input condition for each rule is here listed on the left hand side, with semicolons indicating logical conjunction and slashes logical disjunction. The right hand side lists the resulting suffix. 21

2.6 Hypothetical artificial language. Left table shows training vocabulary, right shows test outputs from four speakers. A regularizes the majority variant, B probability-matches, C regularizes conditional variation, and D regularizes the minority variant. 32

3.1 German plural system with examples, ordered by CELEX type frequency (Sonnenstuhl and Huth, 2002). 55

3.2 Entropy measurements over 6 plural classes $H(C)$ and mutual information between plural class and gender $MI(C;G)$. Compare reported measurements from Williams et al. (2020, Table 3) over a 16-class grouping of the same corpus. 56

3.3 Per-plural-class and overall entropy measurements for final character (F) and grammatical gender G. $|F,G|$: number of observed feature combinations. Entropy calculations are based on the observed frequency of feature values within the lexemes in a particular plural class, e.g. in the top row $H(F,G) = H(F,G \mid \text{-}e), H(F) = H(F \mid \text{-}e)$, etc. 56

3.4 Productive rules of the German plural according to Yang (2016). RFS indicates Reduced Final Syllable, where the singular form ends in *-ər*, *-əl*, or *-ən*. 63

3.5 Plural production data, Köpcke (1988, Table 3). Entropy and percent majority calculations conducted by the author. 71

3.6 Rhyme and Non-Rhyme stimuli used in experiments, originally developed by Marcus et al. (1995). Scores report negative log likelihoods assigned to the stimuli by a character n-gram model (Heafield, 2011) trained on the CELEX noun lexicon. 73

3.7 Marcus et al. (1995) study design. 74

3.8 Marcus et al. (1995) example rating task for *Kach* in the Borrowing condition. 75

4.1 Survey results. Production reported as percentages out of all Rhymes (R) and Non-Rhymes (NR); ratings are averages over a 1 (worst) – 5 (best) scale, with standard errors in parentheses. Highest numbers in each category are bolded. . 92

4.2 Distribution (percentages) of plural class for 1) nouns overall, 2) only neuter nouns, 3) nouns rhyming with M95 stimuli, 4) one-syllable nouns from Unimorph German dataset (Kirov et al., 2016). 93

4.3 Model accuracy (N) by UniMorph corpus split, averaged over 25 random initializations. 95

4.4 Model results by plural suffix for: (left) test set performance (averaged over plural seed); (right) production percentages for rhyme (R) and non-rhyme (NR) M95 stimuli, and correlation (Spearman's ρ) to speaker productions. 95

4.5 Percentage of singular nouns with a given ending which take an *-s* plural, calculated from UniMorph v1 Sylak-Glassman et al. (2015). 106

5.1 Speaker data only: statistical effect of different gender reference levels in contrast coding. Header shows reference level, cells show estimated coefficient (standard error in parentheses). Estimates for Masculine reference level are identical to rows already shown (e.g. suffix *-en*: $-.19$ for gdr.neut, $.31$ for gdr.fem). Stars indicate significance level: * ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001. . 126

5.2 Correlations (Pearson's *r*, 95% confidence intervals in parentheses below) between item-level production percentages for speakers and ED model with 1) overall type frequency (Overall-TF), 2) gender-conditioned type frequency (Gender-TF), 3) each other. 126

5.3 Correlations (Pearson's *r*, 95% CI in parens) between item-level production percentages for speakers and ED model with 1) overall type frequency (Overall-TF), 2) gender-conditioned type frequency (Gender-TF), *only* considering consonant-final monosyllabic nouns in UniMorph (shown in Figure 5.4). 127

5.4 German-language instructions for Study 3 and their English translation. 129

5.5 Mutual information (MI) between gender and plural class for Study 2 and Study 3. The additional manipulation in S3 produces only a small MI increase. 130

5.6 Correlations (Pearson's *r*, 95% CI in parens) between item-level production percentages for speakers and ED model *with* and *without* explicit grammatical gender indicated.
Correlation was evaluated across three bins per item and gender: percent *-e* produced, percent *-en* produced, and percent all other productions. Item-level speaker data is more correlated to the productions of the model without gender than the model with gender. 131

6.1 CELEX2 entropy measurements for gender $H(G)$, plural class $H(C)$, mutual information between plural class and gender $MI(C;G)$, and percentage plural variation explained by gender $U(C|G)$. We see similar values whether using our simplified 3-class analysis or a more traditional 6-class analysis for C. . . . 139

6.2 Speaker strategy classification. Relative to all nouns, most speakers regularize overall plural class variation while *increasing* variability with respect to gender (upper table). Relative to monoCF nouns, most speakers probability-match overall and gender-conditioned plural class variation (lower). 147

6.3 Table 5.5 updated to include S4. Overall mutual information between gender and plural class increases in S4, but remains numerically small. 149

7.1 Calculated values for lexical baselines. 157

7.2 Summary of evaluation data from behavioral experiments. 158

7.3 Summary of models and baselines compared. Baselines are static, while model predictions typically vary for each stimulus item. Each model is trained on either the UniMorph (UM) or Celex (CLX) lexical resource. See text for further description. 159

7.4 Overview of results. For each baseline and model, the mean (standard deviation) and median Jensen-Shannon divergence (JSD) from speaker productions, mean item-level entropy, and mean difference to speaker production entropy for the same item. Speaker entropy values are included for comparison, with item-level differences calculated from the average across all studies. 162

Chapter 1

Introduction

Inflectional morphology has long been a testbed for cognitive theories of linguistic generalization. Most notably, the past tense debate (Rumelhart and McClelland, 1986; Pinker and Prince, 1988) asked whether a statistical model trained on the lexicon could show human-like generalization of morphological categories. In this line of work, many researchers used "wug testing" (Berko, 1958) to assess how speakers inflect unknown words, and then evaluated computational models on their fit to the speakers' behavioral data (e.g. Hahn and Nakisa, 2000; Albright and Hayes, 2003). Despite critical bodies of evidence amassed on both sides, the debate remained largely unresolved (Seidenberg and Plaut, 2014).

In recent years, however, three key research developments have opened up new perspectives on inflection generalization. In linguistics, novel information-theoretic approaches analyze morphological **regularity** in the lexicon in terms of predictability (e.g. Ackerman and Malouf, 2013; Blevins et al., 2017; Wu et al., 2019). In psycholinguistics, a growing literature on **regularization** offers new insights into how speakers learn and generalize linguistic variation (e.g. Hudson Kam and Newport, 2005, 2009; Ferdinand et al., 2019). Finally, **computational models** of natural language have become much more powerful, leading some researchers to claim that artificial neural networks constitute cognitive models of morphological inflection (Kirov and Cotterell, 2018). These methodological innovations share a focus on the statistical *distributions* which characterize inflectional variants. In doing so, they offer new tools to analyze and model morphological generalization, potentially resolving long-standing research questions.

This dissertation adopts the general modeling framework of the past tense debate. The computational task at hand is to learn a statistical model of the lexicon, in the sense of a map from words (inputs) to their inflection classes (outputs), and evaluate how closely its predictions mirror human speaker inflections for those same input words. I focus on German plural inflection, a complex system which has attracted a great deal of attention (c.f. Clahsen, 1999*a*, and responses) due to its particular challenges for statistical models. I use behavioral evidence and modeling to address the question at the heart of the past tense debate: How do human speakers generalize inflectional categories, and can computational models learn to generalize similarly?

1.1 Regularity and German plural inflection

Regularity is a historically important concept in linguistics, yet its meaning is contested and unclear (Herce, 2019). We will consider more precise definitions in Chapter 2, but in the context of the past tense debate, a "regular" inflection class is linguistically *productive* — it can generalize to novel words, although this capacity may be constrained by lexical or other factors. The past tense debate took its name from modeling research focused on English verb inflection. Fortunately, the English past tense is relatively simple, with one inflection class widely recognized as the regular form; in this domain, computational models show human-like behavior mainly by generalizing the regular past tense suffix -*ed* to unknown verbs (Rumelhart and McClelland, 1986; Pinker and Prince, 1988; Albright and Hayes, 2003; Kirov and Cotterell, 2018). As the scope of inquiry expanded to include other inflection systems, researchers continued to place *regularity* at the core of the debate on morphological generalization.

Plural class generalization in German is not quite as straightforward. While the vast majority of English verbs take the regular past tense form, there is no majority class in the German plural system: there are at least five different plural suffixes, and no suffix combines with a majority of nouns in the lexicon. The first generative linguistic analysis of German plural inflection yielded a decision tree with 15 rules and 21 lists of exceptions (Mugdan, 1977). Identifying one or more German plural classes as regular should be difficult for statistical models. Indeed, as we shall see, it has proven quite challenging for linguists.

Within the context of the past tense debate, linguists have proposed three main theoretical accounts of regularity in the German plural system, using different criteria to identify which class or classes are regular. The **minority default** theory (Marcus et al., 1995; Clahsen, 1999*a*) draws on a *rule generation* criterion: the output inflection class which appears in the greatest diversity of linguistic environments must be produced by a *default rule* with minimal restrictions on its input, meaning that this class can productively generalize to the broadest set of input forms. Marcus et al. (1995) identified the rare suffix *-s* as the regular, default plural class, an analysis which remains influential in theoretical linguistic treatments of German (e.g. Trommer, 2020; Schuhmann and Putnam, 2021). By contrast, **schema** theory (Köpcke, 1988; Bybee, 1995; Köpcke et al., 2021) posits that the determinant of productivity is the *type frequency* of the output class; on this view, only plural classes with high type frequency can be considered regular. This frequency criterion reflects a broader view of language as shaped by general social and cognitive processes (Bybee, 2010). Finally, the **gender-conditioned default** theory — arguably the oldest linguistic analysis of German plural inflection (Bloomfield, 1933; Augst, 1979; Bittner, 1994; Indefrey, 1999) — incorporates elements of both former analyses. One can divide the lexicon by gender and identify a conditional majority plural class within each division: *-en* for feminine nouns, and *-e* for nonfeminine. This theory combines a minimal condition on the input — grammatical gender, a property common to all nouns — with output type frequency to identify two plural classes as regular. These three theories represent the main approaches to broadly characterizing regularity in the German noun lexicon.

1.2 Shifting toward distributions

In recent years, some linguists have construed morphological regularity as *predictability*, reflecting any lexical pattern or structure which reduces uncertainty in the prediction of unknown word forms (e.g. Ackerman et al., 2009; Blevins et al., 2017). The Low Conditional Entropy Conjecture (LCEC; Ackerman and Malouf, 2013) exemplifies this approach: "conditional entropies [...] facilitate good guesses in the face of uncertainty" (2013, 436). With respect to German plural inflection, the LCEC predicts that both type frequency and grammatical gender should influence how speakers generalize plural classes, as both of these factors reduce uncertainty. I argue in §3.2.2 that this predictability-based view of morphological regularity aligns most closely with the traditional gender-conditioned default theory in terms of predicted outcomes. It represents, however, a critical shift in perspective: regularity is no longer evaluated in terms of individual inflection classes, but instead characterizes morphological generalization as fundamentally *distributional*.

A similar emphasis on distributions has led to methodological innovations in psycholinguistics. In particular, recent work on **regularization** in artificial language learning (e.g. Hudson Kam and Newport, 2005, 2009) has introduced tools for more fine-grained analysis of how individual speakers generalize morphology. Under this framework, speakers' behavioral patterns are classified as either *probability-matching* to the lexical frequencies of a reference distribution, or *regularizing* — i.e. increasing regularity or predictability relative to the reference. Ferdinand et al. (2019) demonstrate that regularization can also be formalized in information-theoretic terms, connecting this behavior with the functionalist view of the LCEC: speakers produce regularity in the lexicon by increasing the predictability of lexical distributions.

These distribution-focused innovations can reshape our understanding of how speakers generalize German plural classes. In an influential behavioral experiment, Marcus et al. (1995) tested the minority default hypothesis by having adult German speakers rate the acceptability of plural forms for novel nouns. The nouns were either phonologically *typical* and similar to existing words (e.g. "Bral," "Spert"), or phonologically *atypical* due to containing acceptable but rare character sequences (e.g. "Bnaupf," "Fneik"). Speakers preferred the rare plural suffix *-s* for the atypical forms relative to the typical forms. Marcus et al. interpreted this finding in support of their minority default analysis of *-s*: if the regular, default class serves as a 'fallback' for words which do not meet the more restricted membership criteria for other inflection classes, then speakers are predicted to generalize the default class more frequently to unusual words. We can reformulate this prediction in terms of regularization and conditional probability-matching to support more fine-grained behavioral analysis.

Finally, distribution-focused morphological analysis has also been supported by advances in **computational modeling**, especially in the area of sequence generation by artificial neural networks (ANNs). Statistical models of language featured prominently in the past tense debate. Researchers disagreed on whether statistical learning could sufficiently model the linguistic structure underlying how human speakers generalize. One key limitation at the time was that statistical models could not produce open-ended symbolic sequences — they could only classify lexemes based on a pre-selected set of inflection categories (Pinker and Ullman, 2002). This issue has been largely resolved with the advent of neural sequence-to-sequence models, which have also facilitated further analyses of *predictability* in morphological organization (e.g. Cotterell, Kirov, Hulden and Eisner, 2018; Wu et al., 2019; Williams et al., 2020). Kirov and Cotterell (2018) further propose that the sequence-to-sequence capacities of modern ANNs have resolved the past tense debate. They show that a recurrent neural model, the morphological Encoder-Decoder (ED; Kann and Schütze, 2016) achieves speaker-like morphological generalization of the English past tense, although the reliability of this finding has been called into question (Corkery et al., 2019).

Although sequence-generating models resolve one crucial limitation informing the past tense debate, another potential limitation remains: modern ANNs continue to struggle with out-of-distribution generalization (Hupkes et al., 2022), particularly for categories which are infrequent in their training data (Razeghi et al., 2022). This suggests that German plural inflection may remain challenging for modern neural models, as it requires generalization of infrequent classes such as *-s*. Marcus et al. (1995) claimed that ANNs would struggle to capture their key behavioral finding that speakers prefer *-s* for phonologically atypical nouns.

1.3 Summary and research questions

To summarize, German plural inflection has historically been identified as a key challenge area at the intersection of linguistic theory and computational cognitive science, leading to extensive research and debate (c.f. Clahsen, 1999*a*, and responses). Linguists have advanced three broad accounts of how speakers generalize German plural classes: the *minority default* account, which relies upon a **rule generation** analysis to predict the generalization properties of the infrequent class *-s*; the *gender-conditioned default* account, which I claim (§3.2.2) aligns with the **predictability** view in emphasizing conditional relations in the lexicon; and *schema theory*, which highlights the **frequency** of inflection classes within a broader usage-based view of cognition. Despite these theoretical proposals, behavioral experiments have not yet produced a conclusive account of how German speakers generalize plural classes, especially to out-of-distribution or unusual forms; for instance, Marcus et al. (1995) and Zaretsky and Lange (2016) report conflicting behavioral results using the same stimuli. The psycholinguistic literature on regularization provides individual-level behavioral classifications which may address these discrepancies. This leads to the first research question of this dissertation:

- Q1. How do German speakers generalize plural classes to the novel nouns developed by Marcus et al. (1995)? Do speakers show the behavioral patterns predicted by any of the three proposed theories?

Recently, distribution-focused linguistic approaches such as the Low Conditional Entropy Conjecture formulate morphological regularity in terms of **predictability**. This functionalist view of morphological generalization grounds it in speaker behavior rather than lexical structure, and is particularly amenable to computational modeling. If predictability is the key factor influencing how speakers generalize inflection, then computational models should be able to capture this behavior — especially powerful neural models, such as the MED proposed by Kirov and Cotterell (2018). This leads to our second question:

- Q2. Do computational models, especially neural models, trained on the German noun lexicon show speaker-like generalization of plural classes?

Answering these two research questions can help us better understand the nature of linguistic generalization in a statistically complex environment: how speakers assign linguistic categories to novel forms with unfamiliar features, and whether statistical learning alone is sufficient to model this behavior. This research program bears on a range of disciplines, from theoretical linguistics (which analysis of morphological regularity has behavioral and computational support?) to computational linguistics (which computational architecture best approximate how speakers generalize?) and cognitive science more broadly. This dissertation does not propose or adjudicate between theories of cognitive architecture, instead using theoretical accounts and computational models primarily as a lens for empirical analysis; nonetheless, the goal is to inform a broader scientific account of human generalization.

1.4 Present work

This dissertation investigates how human speakers and computational models generalize the famously complex German plural inflection system. I consider the minority default, gender-conditioned default, and schema theories of German plural inflection, and evaluate these theories' predictions in production experiments with German speakers, using the novel noun stimuli developed by Marcus et al.. I train a theoretically-motivated range of statistical learning models — a neural sequence-to-sequence model (the morphological Encoder-Decoder or ED, cf.

Chs. 4, 5, and 7; Kann and Schütze, 2016; Kirov and Cotterell, 2018), a symbolic decision tree (the Abduction of Tolerable Productivity or ATP, cf. Chs. 4 and 7; Yang, 2016; Belth et al., 2021), and a Bayesian classifier (cf. Ch. 7; Bürkner, 2017, 2020) — on the German noun lexicon, and compare their generalization behavior to speaker data. I assess theoretical and computational predictions for three aspects of German plural generalization: distribution of the infrequent plural class *-s* (Ch. 4), influence of grammatical gender (Chs. 5, 6, and 7), and within-item variability (Chs. 4, 6, and 7). To support reproducible scientific analysis, all of the speaker data from the behavioral experiments has been freely released. Some of this data has already been used in subsequent research (Haley, 2020; Beser, 2021; Belth et al., 2021; Dankers et al., 2021; Heitmeier et al., 2021; Rosen, 2022).

With respect to Q1, I find that speaker behavior is most consistent with **probability-matching a phonologically conditioned subset of the lexicon**. Across four experiments, speakers do not show behavior consistent with any default theory. They produce similar plural class distributions for phonologically typical and atypical forms — apart from the *-s* generalization pattern identified by Marcus et al., which my analysis attributes to confounded stimuli. Speaker behavior thus appears at odds with the rule generation view, but also the predictability view: their plural class generalizations are at most mildly sensitive to informative cues such as grammatical gender and word-final phonemes, despite experimental manipulations encouraging attention to informative cues. In general, for each novel lexical item, speakers produce a distribution over plural classes which looks quite similar to the type frequencies of monosyllabic rhyming words in the lexicon; they are relatively insensitive to other prevalent statistical cues in the input, and do not appear to use all available information to reduce generalization uncertainty. This pattern of results appears most consistent with the overall perspective of schema theory, in that the type frequency of plural inflection classes appears to be the main factor influencing speaker generalization.

With respect to Q2, I find that most statistical models heavily condition on grammatical gender, a lexical cue to which speakers are relatively insensitive. The neural and symbolic learners both produce overconfident predictions relative to the speakers' behavioral data. Bayesian models can match the within-item variability of speaker productions, but do not outperform simple lexical baselines. In summary, none of the computational models I investigated appear immediately suitable as cognitive models of morphological generalization for this task; they may require additional structure or inductive biases to attend to the distributional properties which inform how speakers generalize, and disregard properties to which speakers are insensitive.

1.5 Chapter Outline

Chapter 2 presents a novel conceptual analysis of morphological regularity, building on Herce (2019). I identify three proposed criteria for morphological regularity: rule generation, focused on input constraints on the distribution of inflection classes; type frequency, focused on the lexical frequency of inflection classes as realized on their output inflected forms; and predictability, which integrates information sources of all kinds. I use this framework to structure my review of the relevant linguistic, psycholinguistic, and computational literature on morphological generalization. In theoretical linguistics, the rule, frequency, and predictability views of morphological regularity have yielded diverse analyses of the lexicon and its structure. In psycholinguistics, these criteria reflect different behavioral patterns observed in both natural language and artificial language experiments; the latter field in particular has reframed regularity in behavioral terms as *regularization* (Hudson Kam and Newport, 2005, 2009). Finally, I review the computational modeling literature and consider how the three views of regularity align with inductive biases of statistical learners.

Chapter 3 applies the conceptual structure developed in Chapter 2 to review the literature on German plural inflection in particular. I consider the main theoretical proposals — minority default theory, schema theory, and gender-conditioned default theory — and how they relate to the rule generation, frequency, and predictability criteria for morphological regularity. I also review the behavioral and computational literature on German plural generalization.

Chapter 4 investigates minority class generalization, which Marcus et al. (1995) propose as a test case to compare the rule generation and frequency accounts of morphological regularity. In Study 1 (S1), I test whether German speakers use the *-s* plural class to regularize, especially with phonologically atypical nouns, and whether statistical models struggle to capture this behavior, as claimed by Marcus et al. (1995). My speaker data reproduce the key behavioral result from the original paper — the increased use of the *-s* plural class for phonologically unusual nouns — but I show that this is likely driven by stimulus confounds. Due to these confounds, all models also readily capture this effect.

Chapter 5 investigates how grammatical gender affects plural class generalization under controlled conditions, i.e. when provided by the experimenter. In Study 2 (S2), I find that speakers are only mildly sensitive to grammatical gender; in a follow-up study (S3), I find that they remain largely insensitive to gender even with additional financial incentives intended to direct attention to cues to plural class. Neural models, on the other hand, are highly sensitive to grammatical gender, as it is a reliable cue to plural class. This supports the notion that they capture predictability, consistent with the LCEC; however, this measure does not appear to align with speaker behavior.

Chapter 6 reports a fourth behavioral experiment (S4) in which speakers assign both gender and plural class categories, thereby forcing their explicit attention to the joint distribution of these two variables. Even under these conditions, I find that the gender effect is consistent with previous experiments, and does not reach the strength observed in the lexicon as a whole. Instead, it appears that speakers probability-match the distribution of plural classes, the distribution of gender categories, and the mutual information between these two variables to a phonologically restricted subset of the lexicon.

In Chapter 7, I analyze the behavioral data gathered in Studies 1–4, with a particular focus on within-item variability. I evaluate neural, symbolic, Bayesian, and exemplar models trained with and without grammatical gender with respect to how closely their predicted distribution of plural classes for each item matches the distribution (measured in entropy and relative entropy) produced by participants in each of the four experiments. I find that a static phonologically-conditioned lexical baseline outperforms all models. In terms of the regularity criteria, this outcome appears most compatible with the frequency account of morphological generalization, partially compatible with the predictability account, and incompatible with the more prominent rule generation accounts.

Chapter 8 concludes the dissertation and considers limitations and future work.

Chapter 2

Background: Morphological Generalization

In order to characterize and model how German speakers generalize plural inflection, this dissertation draws on research literature from theoretical linguistics, psycholinguistics, and computational modeling. This chapter introduces concepts relevant to morphological generalization and explains their general significance. The following chapter will review how researchers have applied these concepts to the problem of German plural inflection (Ch. 3).

Generalization broadly refers to a learner's capacity to apply learned patterns in a novel environment. Within the domain of inflectional morphology, a learner generalizes correctly by producing the appropriate inflected form (output) of an unknown word (input); for example, an English speaker presented with the novel noun "wug" would be expected to produce the plural form "wugs" (Berko, 1958). Morphological generalization has historically been considered an important research problem at the intersection of linguistic theory, cognitive science, and computational modeling, most famously in the context of the past tense debate (Rumelhart and McClelland, 1986; Pinker and Prince, 1988; Seidenberg and Plaut, 2014).

Researchers generally agree that speakers' morphological generalization must be informed by patterns or structure in the lexicon. They tend to disagree on a) how to characterize the relevant lexical structure (traditionally the domain of linguistics), and b) the mechanistic relation between these patterns and speaker behavior (traditionally the domain of psycholinguistics) — although they often agree that these two domains are inextricably linked. In this section, I will first review linguistic approaches to characterizing structure in the lexicon, broadly known as REGULARITY (§2.1). Then I will consider the psycholinguistic literature, particularly recent findings in the REGULARIZATION framework, which describes how speaker behavior can give rise to lexical structure (§2.2). Finally, I will review computational approaches to modeling morphological generalization (§2.3).

2.1 Regularity: Generalization and the lexicon

2.1.1 Defining morphological regularity

In inflectional morphology, regularity is a property commonly ascribed to words (i.e. lexemes — individual entries in the lexicon) and inflection classes alike. For example, in the case of English past tense verb inflection, the suffix *-ed* is widely recognized as the regular class; commensurately, verbs which combine with *-ed* to form the past tense, e.g. *jump–jumped*, are considered regular verbs. Verbs which take some other past tense form, e.g. *run–ran*, are considered irregular. The general intuition is that regular inflected forms follow recognizable patterns, while irregular forms are idiosyncratic to a greater (e.g. *go–went*) or lesser (e.g. *sleep–slept*) extent. This idiosyncrasy means that, in contradistinction to regular inflection classes such as *-ed*, it is somewhat less common for linguists to identify irregular inflection classes; however, it is certainly possible, as we will see in §2.1.3. The following discussion will consider regularity as a property of inflection classes instead of words, unless otherwise indicated.

Despite the widespread use of regularity as a concept in linguistic analysis, Herce (2019) notes that researchers define the term in diverse and sometimes opposed ways. Herce reviews the linguistics literature and identifies five different criteria used to define regularity. I will discuss these criteria here, and conceptually organize them along two independent dimensions. The first dimension distinguishes whether the criterion can be defined for inflection classes considered in ISOLATION, or if it requires knowledge of some joint DISTRIBUTION over multiple categories. This distinction is primary because it reflects a significant conceptual fault line in linguistic theory. The second dimension relies upon our conceptual understanding that the task of morphological generalization involves mapping from a set of inputs — i.e. lexemes, such as *jump, run, go,* and *sleep* in the preceding paragraph — to a set of outputs — i.e. inflected forms such as *jumped, ran, went, slept*. A regularity criterion may be defined with respect to

	Input	Input-Output	Output
Isolation	Rule generation	Concatenativity	
Distribution		Predictability	Type frequency
Dependent Variable		Productivity	

Table 2.1: Five criteria for regularity identified by Herce (2019), organized by two conceptual dimensions: 1) whether regularity is defined with respect to properties of linguistic categories considered in *isolation* as opposed to some lexical *distribution* over multiple categories, and 2) whether regularity is defined with respect to inputs, outputs, or their joint interaction. Productivity is listed separately, as it is the dependent variable for our purposes.

features of the INPUT, the OUTPUT, or the interaction between them. Table 2.1 depicts this conceptual organization, which I will clarify in the remainder of this section.

Productivity Put simply, productivity *is* morphological generalization: a productive lexical pattern extends to novel items.[1] Productivity is frequently identified with regularity. For our current purposes, productivity is not an explanatory factor; rather, it is the **dependent variable** which we seek to predict based on other proposed criteria. Likely for this reason, "productivity" in the literature appears with many of the diverse senses associated with "regularity," and is similarly challenging to define (Bauer, 2001). In terms of our conceptual scheme, productivity is not defined with respect to the first dimension — productivity can be considered a property of inflection classes in isolation (e.g. one can describe the English past tense suffix *-ed* as productive), or defined as the distribution over lexical patterns which are generalized in some particular context (e.g. in an experimental setting). We can, however, assess the productivity criterion with respect to the second dimension. Productivity is evaluated by observing which inflected forms (outputs) are applied to a certain set of inputs (unknown or novel words), so it is defined in terms of a particular **interaction** or mapping between these two sets.

Rule generation The word "regular" derives from *rēgula*, the Latin word meaning "rule." This etymology informs the colloquial understanding of regularity in inflectional morphology: regular inflected forms can be interpreted as following or adhering to some rule, such as "add *-ed* to a verb to form the past tense." "Interpretable as a rule" is, of course, not the same criterion as "generated by a rule." Defining regularity as rule-based *generation*, rather than description, requires two key steps. The first step is to connect rule application with productivity — if a

---

[1] As Herce points out, this criterion is straightforward to evaluate in psycholinguistic studies using wug tests (Berko, 1958, §2.2.1). Productivity, however, can also be analyzed based on historical data as new words enter the lexicon, so it is not exclusively a behavioral criterion.

speaker uses the English past tense suffix *-ed* to inflect an unknown verb, we might reasonably describe the resulting form as generated or produced through application of the "add *-ed*" rule. The second requirement is a more formal approach to rule definition and application, for example as developed in the generative linguistic research program (e.g. Chomsky and Halle, 1968). A rule in this tradition comprises two parts: an INPUT CONDITION indicating the linguistic context in which the rule is applied, and an associated structural TRANSFORMATION resulting from rule application (c.f. Table 2.5, and discussion in §2.1.3). The rule generation criterion relies upon theoretical linguistic analysis to identify certain inflection classes as regular in any particular domain. In terms of our conceptual framework, this criterion considers linguistic categories in **isolation**. Although rules may interact (e.g. through priority ordering or, outside of the generative framework, through probabilistic application), the regularity of a given inflection class with respect to a particular input context will be defined by one rule; this means it can be evaluated independently of other classes, which are generated by other rules. With respect to the second dimension, rule application in the generative tradition is fully determined by the conditions specified in the **input**, with no influence of the transformation or resulting output form.[2] Note that this characterization applies largely to early work in generative linguistics, as reviewed in §2.1.3, and not necessarily to later developments in the field. For instance, Optimality Theory (Prince and Smolensky, 2004) separates output form generation from constraint evaluation, meaning that regular and irregular forms are generated by the same process; in this approach, generation by rule cannot be a distinctive criterion for morphological regularity.

Concatenativity Many morphological processes considered "regular" involve concatenating affixes to a lexeme, such that the resulting inflected form can be easily segmented. For example, the English word *jumped* is readily decomposed into the root verb *jump* and the regular past tense suffix *-ed*. Some researchers posit a gradient criterion for regularity based on how far the inflection process deviates from concatenation. For example, in the domain of English

---

[2] Separately from rule application, it is true that the structural transformation to yield an output form must be specified on the right-hand side when *defining* a rule; this enables us, for example, to identify the transformation *-ed* as the regular English past tense. However, regularity in this approach is still determined primarily by input conditions. For example, in the analysis in Table 2.5, the same suffix is generated by two separate rules — the rule-based generation, or regularity, of this inflection class depends entirely upon the input conditions. See also §2.1.3 on default classes; the "most regular" inflection class under this criterion is the one with the least restricted input conditions.

past tense inflection, Bybee states that "verbs which use this suffix [-*ed*] and a vowel change (kept, slept, left, etc.) are somewhat irregular, verbs that use only a vowel change (bit, drove, struck) are more irregular, and verbs that have both vowel and consonant changes are the most irregular (thought, taught, went)." (1996, 251–252). With respect to the scheme in Table 2.1, this quote illustrates that concatenativity is defined by the **interaction** between input lexemes and output inflected forms. In terms of the first dimension, concatenativity is a property of inflection classes considered in **isolation**, as it can be assessed for each class independently.

Type frequency The type frequency of an inflection class is defined by how many word types in the lexicon, i.e. lexemes, belong to that class. It is distinct from token frequency, or how often the inflection class appears on word tokens in a corpus. Like rule generation, type frequency is intuitively connected to productivity: a productive lexical pattern generalizes to novel items, so its type frequency will naturally increase as new words enter the lexicon. For this reason, some researchers argue that regular inflection classes must have high type frequency relative to other classes (e.g. Blevins et al., 2017). Clearly, this criterion can only be defined with respect to the frequency **distribution** over all inflection classes in a lexicon. For our second dimension, only **output** inflected forms are relevant to calculating type frequency.[3]

Predictability Herce (2019) notes that many researchers describe regular inflection as "predictable" in some informal sense; for instance, Pinker and Ullman state that the English regular past tense suffix "applies predictably to thousands of verbs" (2002, 457). For the most part, however, "predictability" has not been sufficiently formalized to support detailed theoretical or empirical analysis. This issue has been addressed in recent work. In particular, Ackerman and Malouf (2013) propose an information-theoretic characterization of complexity in inflectional morphology, which Cotterell and colleagues have advanced as a criterion for morphological regularity (Cotterell, Kirov, Hulden and Eisner, 2018; Wu et al., 2019). For the purposes of this dissertation, I take Ackerman and Malouf's Low Conditional Entropy Conjecture (LCEC; §2.1.2) to represent a formal predictability criterion for regularity. This approach is also closely tied to generalization and productivity: under a predictability account, the core function of regularity in the lexicon is to enable prediction for unknown inflected forms. With respect to

---

[3] While the interaction between input lexemes and output forms is relevant in *defining* inflection classes, only type frequency in the output used to determine *regularity* under this criterion.

our conceptual scheme (Table 2.1), an information theoretic criterion is by definition characterized in **distributional** terms. Furthermore, Ackerman and Malouf clarify that predictability[4] is facilitated by "[a]ny information that helps speakers predict the realization of the word" (2013, 439). This includes features of the input (e.g. phonological form, grammatical gender) as well as the output (e.g. type frequency), so this criterion is defined by the **interaction** of these two sets of wordforms.

Summary Although the regularity criteria reviewed here can be considered logically separate, as we have seen, they all share some degree of conceptual relation. For example, rule generation implies productivity, and productivity implies type frequency. Herce (2019) observes that, in practice, most linguistic analyses draw on multiple criteria in assessing morphological regularity. In Chapter 3, I review how these criteria have informed different theoretical approaches to characterizing regularity in the domain of German plural inflection.

The different criteria proposed for morphological regularity reflect conflicting theories of how to properly characterize patterns or structure in the lexicon. Researchers often (though not always) further assume that lexical structure directly influences how speakers generalize morphological categories, and therefore that behavioral evidence from speakers can help distinguish between competing accounts of lexical structure. To that end, we will review psycholinguistic approaches to studying morphological generalization (§2.2). Different conceptions of lexical structure have also informed various computational approaches to modeling morphological generalization and the lexicon (§2.3).

Before that, however, we will consider how three key criteria have informed competing linguistic approaches to characterizing lexical structure. First, we will formalize **predictability** in terms of information theory, and consider how these concepts relate to morphological structure under the assumptions of the Low Conditional Entropy Conjecture (§2.1.2). Then we will consider the theoretical linguistic concept of a DEFAULT, and how it has shaped the **rule generation** view of morphological regularity (§2.1.3). Finally, we briefly review the literature on **type frequency** and lexical structure, and discuss how this criterion interacts with both the rule generation and predictability views (§2.1.4).

---

[4] "Predictability" here specifically means "entropy reduction;" see discussion in §2.1.2.

		SG.INDEF	SG.DEF	PL.INDEF	PL.DEF
C1	'school'	*skol**a***	*skol**an***	*skol**or***	*skol**orna***
	'bottle'	*flask**a***	*flask**an***	*flask**or***	*flask**orna***
C2	'chair'	*stol*	*stol**en***	*stol**ar***	*stol**arna***
	'box'	*ask*	*ask**en***	*ask**ar***	*ask**arna***
C3	'idol'	*idol*	*idol**en***	*idol**er***	*idol**erna***
	'Basque'	*bask*	*bask**en***	*bask**er***	*bask**erna***

Table 2.2: Paradigms illustrating three inflectional classes (C1–C3) in Swedish nouns, reproduced from Round et al. (2022). Morphological exponents are indicated with boldface.

	SG.INDEF	SG.DEF	PL.INDEF	PL.DEF
C1	*-a*	*-an*	*-or*	*-orna*
C2	*ø*	*-en*	*-ar*	*-arna*
C3	*ø*	*-en*	*-er*	*-erna*

Table 2.3: Arrangement of Swedish nominal exponents from Table 2.2, according to inflection class (row) and grammatical feature (column).

2.1.2 Information theory and the Low Conditional Entropy Conjecture

The Low Conditional Entropy Conjecture (LCEC; Ackerman and Malouf, 2013) posits that morphological generalization — the need for speakers to produce unknown inflected forms — is the core functional pressure organizing the lexicon. The resulting lexical structure is realized in the form of *predictive relationships* between members of morphological PARADIGMS. Ackerman and Malouf use 'paradigm' to mean the particular arrangement of morphological EXPONENTS associated with some linguistic domain over all grammatical features and inflection classes, for instance in Table 2.3. Here, I follow the terminology of researchers who work with wordforms rather than exponents (e.g. Cotterell, Kirov, Hulden and Eisner, 2018; Round et al., 2022) in using 'paradigm' to mean the set of wordforms associated with a particular lexeme — i.e. each row in Table 2.2 constitutes a separate paradigm — and 'inflection class' to mean a particular set of morphological exponents, and/or the set of lexemes which take those exponents.

Morphological generalization is straightforward when only one inflection class exists, but can be challenging when multiple inflection classes are present. Ackerman and Malouf propose two different information-theoretic measures to quantify morphological complexity. The first metric, E-COMPLEXITY (where E stands for Enumerative), reflects the difficulty of guessing the inflection class of an unknown lexeme, absent any other information. The second metric, I-COMPLEXITY (where I stands for Integrative), re-estimates the difficulty of inflection class assignment in the presence of extra information — for instance, conditional on having observed another wordform in the same paradigm.

Entropy and E-complexity Ackerman and Malouf use Shannon entropy (Shannon, 1948) to quantify E-complexity. Given a category C with possible values $c \in C$, the entropy in bits $H(C)$ can be defined in terms of the probability distribution $P(C)$:

$$H(C) = -\sum_{c \in C} \mathrm{P}(c) \log_2 \mathrm{P}(c) \tag{2.1}$$

Ackerman and Malouf consider three different approaches to quantifying E-complexity, which I will illustrate here in relation to Table 2.3. The first, DECLENSION ENTROPY $H(D)$, assumes that each inflection class is equally probable. Given three inflection classes for Swedish nouns, this would work out to roughly 1.58 bits:

$$H(D) = -\sum^{3} \frac{1}{3} \log_2 (\frac{1}{3}) \approx 1.58 \tag{2.2}$$

The second approach, $H_{TF}(D)$, weights the probability of each inflection class based on its *type frequency* (TF), or how many lexemes belong to that class. This measure is strictly equal to or lower than $H(D)$ by definition, as uniform probability over each class produces the maximal entropy value for a given distribution.

The third approach, PARADIGM CELL ENTROPY $H(c)$, rescopes the task from predicting the inflection class of an unknown lexeme to predicting its realization in a specific paradigm cell (once more assuming uniform probability over inflection classes). This measure is motivated by the insight that, during real world language use, speakers typically have to generalize only one inflected form for a particular grammatical feature combination, which is an easier task than producing the full paradigm for a given lexeme. For example, in the SG.DEF column in Table 2.3, we see that both C2 and C3 are realized with the suffix *-en*, so an inflected form with this exponent has probability $\frac{2}{3}$ relative to $\frac{1}{3}$ for *-an*:

$$H(\textsc{sg.def}) = \frac{2}{3} \log_2 (\frac{2}{3}) + \frac{1}{3} \log_2 (\frac{1}{3}) \approx 0.92 \tag{2.3}$$

To obtain the paradigm cell entropy over an entire system, we calculate $H(c)$ for each cell c and then take the average. This measure is always strictly lower than $H(D)$, as the number of possible exponents for a given paradigm cell will always be less than or equal to the number of inflection classes. The average paradigm cell entropy for Table 2.3 is 1.25 bits.

All three of these approaches estimate E-complexity, and more approaches are certainly conceivable; for example, paradigm cell entropy could be weighted by type frequency as well. Johnson et al. (2020) use average paradigm cell entropy as presented here to measure E-complexity in an artificial language learning study. Cotterell, Kirov, Hulden and Eisner (2018) propose a different method to estimate E-complexity using generative models of wordforms rather than exponents, discussed further in §2.3.3. Note that, in terms of the conceptual scheme shown in Table 2.1, all of the E-complexity measures considered here are defined exclusively with respect to distributions over outputs (i.e. inflected forms).

Conditional entropy and I-complexity While E-complexity quantifies the challenge of morphological generalization in terms of outputs, Ackerman and Malouf argue that this overstates the problem: speakers actually face a much easier task due to *implicative relationships* between paradigm cells, which is measured by I-complexity. I-complexity relies upon conditional entropy to quantify the predictability of a specific unseen inflected form, conditional on having observed another form from the same lexeme, i.e. another wordform in the same paradigm. Given a category C_1 with possible values $c_1 \in C_1$, its conditional entropy with respect to another category C_2 can be defined (inter alia) in terms of the probability distribution $P(C_2)$ and the conditional probability distribution $P(C_1 \mid C_2)$:

$$H(C_1 \mid C_2) = -\sum_{c_2 \in C_2} \mathrm{P}(c_2) \sum_{c_1 \in C_1} \mathrm{P}(c_1 \mid c_2) \log_2 \mathrm{P}(c_1 \mid c_2) \tag{2.4}$$

As for E-complexity, many approaches can be used to estimate the relevant distributions. Ackerman and Malouf calculate I-complexity in terms of the AVERAGE CONDITIONAL ENTROPY between pairs of cells, again assuming uniform probability over inflection classes. For example, in Table 2.3, $H(\text{SG.DEF} \mid \text{SG.INDEF}) = 0$, because the SG.DEF form of a lexeme is perfectly predictable given knowledge of its SG.INDEF form. On the other hand, $H(\text{PL.INDEF} \mid \text{SG.INDEF}) = \frac{2}{3}$, because $H(\text{PL.INDEF} \mid \text{SG.INDEF} = \text{-}a) = 0$ (class C1, probability $\frac{1}{3}$) and $H(\text{PL.INDEF} \mid \text{SG.INDEF} = \emptyset) = 1$ (classes C2 and C3, probability $\frac{2}{3}$). To estimate the I-complexity of the overall system, we repeat the calculation over all pairs of cells and average the result. In this case, we obtain an I-complexity of $\frac{1}{3}$ bits for the Swedish nominal system shown in Table 2.3, substantially lower than the E-complexity measure of 1.25 bits.

Ackerman and Malouf (2013) argue that typologically diverse languages with a wide range of values for E-complexity nonetheless share relatively low I-complexity. Their analysis supports the Low Conditional Entropy Conjecture that low I-complexity is the key factor enabling speakers to generalize lexical structure to unknown words within complex inflectional systems. This approach — namely, characterizing regularity as predictability in information-theoretic terms — can flexibly integrate various sources of distributional information, from features of input lexemes (e.g. observed from another paradigm cell) to frequencies of output forms. This motivates the classification of this approach in Table 2.1.

Chapter 3 reviews various information-theoretic properties of the German plural system (§3.1). To the best of my knowledge, the LCEC analysis described here has not been applied to German plural inflection.[5] Due to the high degree of mutual information (i.e. conditional entropy reduction) between grammatical gender and plural class (Table 3.2), I argue that in practice, the LCEC/predictability account most closely aligns with a more traditional generative linguistic analysis that assigns default plural classes based on grammatical gender (§3.2.2). In the following section, I review the theoretical literature on defaults and morphological regularity.

2.1.3 Defaults in linguistic theory

This section introduces the theoretical linguistic concept of a DEFAULT category. I first discuss the background and significance of defaults in linguistic theory more broadly, then introduce the related concept of the Elsewhere Distribution. Finally, I consider how linguistic defaults relate to morphological regularity, in particular the rule generation and predictability criteria.

Defining defaults The term DEFAULT has a range of meanings and uses in linguistics, but they largely express the same principle: for some given linguistic domain, the default is the *general case*, which may be overridden by more specific cases — but if more specific conditions do not apply, we should expect the default (e.g. Zwicky, 1986; Gisborne and Hippisley, 2017). This principle has a long history in linguistic analysis, first appearing around 500 BC in the Sanskrit analysis of pioneering grammarian Pāṇini (Deo, 2007). Kiparsky (1973) reintroduced Pāṇini's principle to address rule interaction in the context of modern generative linguistics.

---

[5] Cotterell, Kirov, Hulden and Eisner (2018), discussed in §2.3.3, estimate the E- and I-complexity of the German nominal system as a whole, which goes beyond the nominal plural subsystem of interest here.

	SINGULAR			PLURAL
	masculine	neuter	feminine	
nominative	*-e*	*-e*	*-e*	*-(e)n*
accusative	*-(e)n*	*-e*	*-e*	*-(e)n*
dative	*-(e)n*	*-(e)n*	*-(e)n*	*-(e)n*
genitive	*-(e)n*	*-(e)n*	*-(e)n*	*-(e)n*

Table 2.4: Suffixes for German weak adjective inflection. German adjectives are inflected to agree with nouns in case, grammatical gender, and number.

R1. [+ACC;+MASC;+SG] → *-(e)n*
R2. [+NOM/ACC;+SG] → *-e*
R3. [] → *-(e)n*

Table 2.5: Generative analysis of German weak adjective inflection (Zwicky, 1985). The input condition for each rule is here listed on the left hand side, with semicolons indicating logical conjunction and slashes logical disjunction. The right hand side lists the resulting suffix.

As discussed earlier, a rule in the generative linguistic tradition comprises an INPUT CONDITION indicating the linguistic context in which the rule is applied, and an associated structural TRANSFORMATION resulting from rule application. The input condition determines the rule's scope, and this scope in turn determines which rules are prioritized: "when two rules conflict, the more specific takes precedence over the more general" (Anderson, 1982, 593). Note that this exclusive focus on the *input* to a morphological process is consistent with the rule-based characterization of morphological regularity presented in Table 2.1.

Zwicky's (1985) analysis of German weak[6] adjective inflection illustrates how the default principle affects rule ordering in generative linguistics (also discussed by Brown, 2016; Gisborne and Hippisley, 2017). Table 2.4 shows the distribution of suffixes taken by adjectives during weak inflection, based on different combinations of the grammatical features case, gender, and number. To account for the observed set of suffixes, Zwicky posits three rules, which we present in a simplified format in 2.5. For instance, only adjectives agreeing with masculine, accusative, singular nouns would match the input condition for R1, while adjectives agreeing with nominative or accusative singular nouns would match R2. By listing the most conjunctive features, R1 selects a proper subset of the linguistic forms selected by R2 and is thus more specific. R3 does not specify any features in its input condition, making it the most general rule of all. Brown describes Zwicky's analysis as "an implicit layering of defaults [...] one very specific rule, one more general, and one even more general" (2016, 277). As discussed below,

---

[6] "Weak" adjective inflection occurs in German when the noun phrase is preceded by a determiner. "Strong" inflection occurs when the noun phrase has no determiner, and follows a different pattern.

only the rule with an unrestricted input condition — here, R3 — technically qualifies as the default rule, but Brown's observation shows how this concept reflects the broader linguistic principle: more specific cases take precedence at each level of generality, where specificity is evaluated with respect to the feature values posited in the input conditions.

The Elsewhere Distribution Kiparsky (1973) introduces the term ELSEWHERE CONDITION to refer to negatively-defined linguistic contexts in phonology, later extending to morphology (Kiparsky, 1982). A default rule must have the Elsewhere Condition as its input criterion (e.g. Zwicky, 1986). For example, in Table 2.5, the left hand side of R3 represents the Elsewhere Condition. It does not specify any constraining features — within the domain of weak adjective inflection, R3 applies *elsewhere*, which is to say in any context where the more specific R1 or R2 does not apply. This lack of environmental restrictions is the default rule's defining feature: "a default rule is seen as being context-free in the sense that it applies in the absence of any other information" (Brown, 2016, 280).

The Elsewhere Condition is a definitional criterion for a default *rule*; as such, it must be associated with a particular theoretical analysis, i.e. set of rules. How do linguists find evidence to support such analyses? In general, default rule analyses depend upon the linking assumption that the Elsewhere Condition is realized by its observational counterpart, the ELSEWHERE DISTRIBUTION. The Elsewhere Distribution is identified by analyzing how outputs,[7] i.e. inflected forms, are distributed with respect to features of the input. A linguistic variant which appears in a diverse range of heterogeneous contexts is said to be *elsewhere distributed*; its distribution cannot be easily summarized with reference to specific properties of the input, so the linguist infers that this variant follows the Elsewhere Condition, i.e. is assigned by a default rule. For example, the data presented in Table 2.4 shows that the suffix *-(e)n* appears with the widest variety of feature combinations, making it the inflectional class which most plausibly reflects the Elsewhere Distribution. Even though another hypothetical analysis of the same data may propose different rules from from Table 2.5, if the analysis were to include a default rule, that rule would almost certainly assign *-(e)n* rather than *-e*.

---

[7] This may appear to contrast with the input focus that I claim characterizes rule-based approaches to morphological organization; however, as explained here, the main theoretical import of this distribution is to identify the output class which is associated with the Elsewhere Condition in the input.

Defaults and regularity As discussed in §2.1.1, proposed criteria for "regularity" are most readily understood as attempts to account for productivity, i.e. morphological generalization. The rule generation criterion posits that a regular inflection class is one that is generated by rule, under the assumption that the rule will productively apply to any novel input lexemes which match the features specified in its input condition. This approach, however, raises questions for inflection systems which are analyzed in terms of multiple rules. For example, consider the analysis in Table 2.5, which posits that all German weak adjective inflection classes are rule-generated. Does this mean that both *-e* and *-(e)n* are "regular" classes in this system?

The notion of a "default" helps to resolve this problem: if regularity, and hence morphological generalization, is best characterized with respect to the input conditions of productive rules, then the most productive rule should be the one with the least restrictive conditions, as it can generalize to the largest set of potential input lexemes. This reasoning implies that rule ordering according to Pāṇini's principle also defines a hierarchy[8] in terms of regularity, with the default inflection class qualifying as "most regular" due to its potential to generalize to any input. The Elsewhere Distribution of the default class, then, renders it indisputably regular. Other inflection classes with more restricted scopes (such as R2 in Table 2.5) might be produced by "irregular rules" (Yang, 2017) or "sub-regular rules" (Royle et al., 2012) — phrases which are uninterpretable if "regular" simply means "rule-generated." The default class, however, can apply to any input, making it regular by definition.

Though this understanding of regularity is not formulated in information-theoretic terms, we can sketch out some intuitive connections between the default rule generation criterion and principles of statistical learning. The default class is the one which appears with the most heterogeneous feature combinations in the input space. In other words, the default is the output class associated with the *highest entropy in the input space*, i.e. the most variable distribution of probability mass over linguistic environments. We can view the Elsewhere Condition as a consideration of grammatical *economy* (Wilder et al., 1996) related to minimizing *description length*: the variable distribution of input features associated with the default class is most efficiently characterized in negative terms (Wilder and Gärtner, 1997). Indeed, Rasin et al.

---

[8] The importance of hierarchical rule ordering somewhat weakens my claim that regularity in the rule generation approach can be defined in isolation (§2.1.1), as in practice, most analyses with default rules clearly rely upon the presence of earlier pre-emptive rules to explain particular linguistic data. Nonetheless, I maintain that this regularity criterion *can* be evaluated in isolation: any default rule specified by the Elsewhere Condition will identify a default class, and that class qualifies as 'regular' without regard to any rules which may or may not precede it.

(2021) show that the formal principle of Minimum Description Length (MDL; e.g. Rissanen, 1983) encapsulates various grammatical economy evaluation metrics proposed in the generative linguistics literature. MDL, in turn, is closely related to the statistical learning objective of maximizing entropy over a distribution subject to specified constraints (Feder, 1986; Grünwald, 2007) — exactly the aim of the Elsewhere Condition.

To recap, while the generative default rule analysis does not explicitly formulate regularity as the minimization of entropy in prediction (contra Ackerman and Malouf, 2013, §2.1.2), it at least partly aligns with other approaches to statistical learning (e.g. Minimum Description Length) by maximizing entropy in the input space. In Chapter 3, we will consider how this input variability criterion informs default theories of German plural inflection. The minority default theory (§3.2.1) hinges entirely upon this criterion, whereas gender-based default theory (§3.2.2) is less reliant given its alignment with the predictability criterion discussed above (§2.1.2). Crucially, the default output category is defined and evaluated only with respect to the distribution of features in the input space: there is no influence of the relative frequency of different output inflected forms. In the following section, I briefly discuss the third regularity criterion of type frequency, which stands in opposition to the input-focused default rule generation criterion developed in this section.

2.1.4 Type frequency

The final criterion focuses on the type frequency of inflection classes observed upon output inflected forms. This definition of morphological regularity is often invoked informally — for example, Blevins et al. define "regular items" as "members of classes with a high type frequency" (2017, 141) — but has not received the same amount of theoretical consideration or formalization compared to the rule generation and predictability criteria discussed above. For instance, Joan Bybee has contributed many comprehensive research treatments of frequency effects across various levels of linguistic analysis (e.g. Bybee and Slobin, 1982; Bybee, 1995, 1996, 2006*a*,*b*, 2010); even in her work, however, the precise role of frequency remains vague. Bybee notes that "frequency interacts with other factors, such as phonological and semantic similarity, categorization, and semantic/pragmatic change. It is often difficult to discern which factors are the most important in determining linguistic behavior" (2006*a*, 17). This illustrates a barrier to the development of formal linguistic models relying on type frequency: in practice, the type frequency of output inflection classes shows complex patterns of interaction with other linguistic factors. These multifaceted interactions in turn make it unclear when type frequency has a causal role:

> [T]he answer to the question of whether frequency is a cause or an effect is complex. On the one hand, frequency is just a tally, a pattern observable in texts, which is of course an effect. On the other hand, frequency or repetition of experiences has an impact on cognitive representations and in this way becomes a cause [...] (Bybee, 2006*a*, 18)

If type frequency can be qualified through interaction with other linguistic factors, then this criterion may start to look more like predictability (§2.1.2) than an independent dimension of regularity. Can we reliably differentiate these views? Haspelmath (2021) argues that, while frequency asymmetries in linguistic variants may *result* from many diverse factors, they are universally *causal* with respect to predictability. He illustrates this with reference to coding asymmetries, in which more frequent linguistic variants (e.g. present-tense “go”) are expressed with shorter forms compared to less frequent variants (e.g. future-tense “will go”):

> If context is kept constant, higher-frequency meanings are more predictable than lower-frequency meanings because of their frequency: It is less surprising if my interlocutor uses a present-tense form than if she uses a future-tense form [...] Thus, as was already noted in Section 1, the causal chain goes from high frequency to predictability, and from predictability to short coding. (Haspelmath, 2021, 624)

Within morphology, then, the type frequency criterion posits that the frequency of an inflection class over output inflected forms independently contributes to its regularity (i.e. propensity to generalize; §2.1.1), although its generalization may also be influenced by other factors (e.g. linguistic features of the input, relevant sentence context, and so on). To editorialize somewhat, my impression is that advocates of the type frequency criterion show an admirable willingness to acknowledge and explore this complexity, but this unfortunately comes at the expense of generating clear and testable theoretical predictions. This trade-off is evident in the debate on German plural inflection, where the type frequency criterion is advanced within the framework of schema theory (Köpcke, 1988; Bybee, 1995). As discussed in §3.2.3, while the schema theory account of German plural inflection captures key empirical findings, it does not appear sufficiently formalized to generate fine-grained predictions.

Summary In this section, I have reviewed different criteria which have been proposed to characterize *regularity*, or patterned structure in the lexicon informing morphological generalization (§2.1.1). I consider two approaches in detail: the information-theoretic account of regularity as **predictability**, exemplified by Ackerman and Malouf's Low Conditional Entropy Conjecture (§2.1.2), and the generative **rule generation** account of regularity, under which the most productive class in an inflection system is generated by a negatively-conditioned default rule (§2.1.3). I also briefly review the **type frequency** criterion for regularity, which has not received as much theoretical development (§2.1.4). The conceptual scheme depicted in Table 2.1 captures key differences between these approaches. The rule generation account focuses on feature-valued input conditions for linguistic categories considered in isolation, while the type frequency account highlights distributions over output inflected forms, and the predictability account integrates distributional information from input lexemes and output forms. All three criteria, however, share the same goal of explaining morphological generalization. Chapter 3 will discuss how these criteria have informed competing linguistic theories of the lexical structure and generalization of German plural inflection.

2.2 Regularization: Generalization and speakers

While the previous section reviewed different approaches to characterizing linguistic structure within the lexicon, this section focuses on the psycholinguistic question of how those lexical patterns affect speaker behavior. These two topics are separable in principle, reflecting the disciplinary divide between linguistic and psycholinguistic research; however, as we have seen, linguistic theories of regularity often seek to account for morphological generalization, which in practice must be realized by speaker behavior at some level of analysis. Given this connection, many linguistic researchers consider behavioral evidence from psycholinguistic experiments essential to meaningfully adjudicate between competing theories of lexical structure. In §2.2.1, I review some general findings from psycholinguistic studies of morphological generalization in relation to three different conceptions of linguistic regularity, namely rule generation, predictability, and type frequency (c.f. Table 2.1).

In recent years, some researchers have inverted the question: instead of asking how regularity in the lexicon influences speaker behavior, they use artificial language learning experiments to investigate how speaker behavior can *produce* lexical regularity, a process known as REGULARIZATION (Hudson Kam and Newport, 2005, 2009). The regularization literature provides fine-grained analytical tools to evaluate the relation between speaker behavior and lexical distributions, for instance in information-theoretic terms (Ferdinand et al., 2019), which intuitively connect to linguistic treatments of regularity as predictability (§2.1.2). In §2.2.2, I introduce key concepts and behavioral findings from this literature, and consider how they relate to morphological generalization in natural language.

2.2.1 Regularity and speaker generalization: Wug tests

The wug test is an experimental design in which speakers are asked to produce or evaluate inflected forms of nonexistent words. For many years, this experimental approach has been the dominant method to assess how speakers generalize inflectional morphology based on the lexical patterns of their language. In addition to morphological generalization, wug tests are often used to investigate how speakers generalize phonological patterns; in the original wug test, Berko (1958) studied phonologically-conditioned alternations of the English plural suffix *-s*, which is deterministically realized as /-s/, /-z/, or /-iz/ based on the final phoneme of the input (as in the words "cats," "dogs," and "glasses" respectively). Wug tests are generally used to characterize speakers' linguistic knowledge of their lexicon, particularly the aspects which influence generalization.

One methodological concern for experimenters conducting wug tests is whether production or evaluation tasks are better suited to address their research questions. Morphological generalization is often framed as a problem for speakers, who must produce inflected forms for unknown words (Ackerman et al., 2009; Ackerman and Malouf, 2013; Blevins et al., 2017). This suggests that production tasks may be a more natural approximation of the real-world problem. On the other hand, evaluation tasks such as acceptability judgments enable gradient responses, which may capture more nuanced influences, or processes relevant to linguistic comprehension. Schütze (2005) considers the wug test from the vantage point of the experimental participants, and distinguishes two possible scenarios: participants may treat a novel stimulus either as an obscure word that already exists in the language (e.g. a rare word one may find in the dictionary), or as a word newly entering the language (e.g. a neologism, or borrowing from another language). Based on an analysis of data from Albright and Hayes (2003), he conjectures that participants may prefer the former (dictionary) interpretation in the context of rating tasks, but prefer the latter (neologism/borrowing) interpretation in production tasks; see also Kawahara 2015 for an empirical comparison of forced-choice production and rating tasks. Based on the consistent divergence in results from production and evaluation tasks, Clahsen (2016) recommends using both methods when possible. This recommendation informs the behavioral study in Chapter 4, which collects production and rating data for the same stimuli from the same speakers.

Rule generation Numerous psycholinguistic studies of morphological generalization have been informed by the assumption that linguistic regularity reflects rule application. Under this view, the input constraints associated with individual inflection classes are the key factor influencing how speakers assign those inflection class to novel words:

> "The generalization properties of inflectional and other morphological processes provide a crucial diagnostic for how they are mentally represented. Some morphological processes may be freely applied to novel or unusual words, others only under appropriate circumstances." (Clahsen, 2016, 798-799)

A process which freely applies to all words, including unusual words, is understood as the default inflection class (c.f. §2.1.3). Clahsen reviews the experimental literature from this perspective, and cites several behavioral[9] studies (e.g. Veríssimo and Clahsen, 2014) which find support for "a variable-based mechanism that generalizes by default to all members of a given grammatical category" (2016, 803).

Researchers have used a range of rule-based models to analyze behavioral data from wug tests. Here, I will highlight three studies of particular relevance to this dissertation. Marcus et al. (1995) conduct a rating experiment with German speakers and interpret their results in terms of a *dual-route model* combining analogical generalization with a rule-generated default class (see also Clahsen, 1999*b*; Pinker and Ullman, 2002, and extensive discussion in §3.2.1).

---

[9] Note that this assumption is not shared by all morphologists working within default frameworks. For instance, Gisborne (2017) and Kihm (2017) argue that morphological default classes have cognitive relevance, but they support their claims with diachronic evidence from generalization through historical change rather than synchronic evidence from behavioral experiments.

In this analysis, only one inflection class is rule-generated — therefore regular — and also the default. Other approaches, however, posit that speaker generalization behavior may be influenced by interacting systems of multiple productive rules. Albright and Hayes (2003) use simultaneous probabilistic application of multiple rules at varying levels of generality to model behavioral results from an experiment on English past tense inflection. Yang (2016) proposes a model of morphological generalization based on serial rule application, which Belth et al. (2021) apply to behavioral wug test data from German and English. Both of these rule-based approaches have been implemented as computational models, and will be discussed in greater detail in §2.3.1. These three studies posit different mechanisms connecting lexical structure — characterized here as rules, as well as analogy in the dual-route case — to speaker behavior, but all share the hypothesis core to the rule generation conception of regularity: that the input conditions associated with individual inflection classes have significant causal effects on how speakers generalize inflection. In later chapters (e.g. Chs. 4 and 7), I evaluate this hypothesis using behavioral wug test data and computational models, including the symbolic learner proposed by Belth et al. (2021).

Type frequency Other studies, however, have found that the frequency distribution over inflection classes in the lexicon may have the strongest influence on how speakers generalize inflection. This suggests that speakers may PROBABILITY-MATCH to the lexical frequencies of inflected forms, i.e. outputs. Pierrehumbert (2022) reviews the behavioral evidence for probability-matching, which has been found mainly in wug tests of phonological patterns, but is also attested for inflectional morphology (e.g. Ernestus and Baayen, 2003; Hayes et al., 2009). She notes that confusion around the topic makes it difficult to find evidence for or against probability-matching in the existing behavioral literature:

> Unfortunately, many of the earlier papers that are cited argue for the importance of probabilities, but not for probability-matching per se. [...] To show that learning is probability-matching, it is necessary to show that the statistical patterns in the output of individual learners match those in the input, which are assumed to be the same as those in the ambient language. [...] One issue is that many studies only report data that has been pooled across participants [...] pooled data can give a spurious appearance of probability-matching in cases where different participants

> learn categorically different systems. (Pierrehumbert, 2022, 647-648)

This gap in the literature regarding morphological generalization of natural language is partly filled by artificial language learning research, where the concept of probability-matching has been more extensively developed and contextualized in the broader cognitive literature (§2.2.2). In Chapter 6, I draw on artificial language learning methodology (e.g. Ferdinand et al., 2019) to evaluate probability-matching behavior on the level of individual speakers. Computational models of probability-matching will be discussed in §2.3.2.

Predictability The predictability criterion for linguistic regularity posits that lexical structure exists precisely in order to facilitate generalization by speakers; this approach is fundamentally motivated by speaker behavior. The Low Conditional Entropy Conjecture (LCEC) has been assessed in artificial language studies, which we will discuss in §2.2.2. Predicting morphological generalization in natural language, however, depends upon accurately characterizing the lexical distributions to which speakers are exposed. Blevins et al. (2017) draw attention to the Zipfian nature of linguistic distributions, and argue that most speakers in fact never encounter the majority of inflected forms in their language. This implies that speakers do not have access to the full lexicon in practice, and must base their predictions for unobserved forms on partially observed morphological paradigms. Blevins et al. claim that speakers overcome this challenge to predict unseen forms via analogical generalization based on LEXICAL NEIGHBORHOODS, i.e. existing words which are phonologically similar to the novel word, and cite Milin et al. (2011) for supporting behavioral evidence from wug tests with Serbian speakers. See also Nieder et al. (2020) for an analogical account of speaker plural generalization in Maltese.

One key challenge for prediction-based approaches to speaker behavior is accounting for patterns in the lexicon which are statistically robust, yet do *not* inform how speakers generalize. Becker et al. (2011) identify a phonological alternation in Turkish which reliably correlates with three factors in the lexicon, but find that speakers only condition on two of those three factors during wug tests. Dawdy-Hesterberg (2014) implements an analogical model of inflection generalization for Arabic verbs. While the model predicts generalization conditional on vocalic indicators of verb patterns, she finds that speakers do not use this information in wug tests, and instead probability-match to a lexical distribution conditioned on a separate morphological template. Gagliardi and colleagues have reported similar statistical insensitivity in first-language acquisition of Tsez noun classes (Gagliardi and Lidz, 2014; Gagliardi et al., 2017). These findings suggest that speakers' morphological generalization is not wholly determined by the predictability of lexical structure; independent cognitive or linguistic biases may render speakers sensitive to certain statistical properties of the lexicon, and insensitive to others. In this dissertation, the question of *which* lexical statistics inform speaker generalization of German plural inflection will be considered in depth in Chapters 5, 6, and 7.

Summary This section has discussed the wug test (Berko, 1958), a psycholinguistic experimental technique to evaluate how speakers generalize inflectional morphology. Wug tests have often been used to provide behavioral evidence for or against different theories of morphological regularity, such as the rule generation, type frequency, and predictability accounts reviewed here. Although the task of inflecting unknown words may appear simple, the complexity of natural language makes result interpretation difficult. Any natural language lexicon contains many different types of organization at different levels, any of which may or may not influence speaker behavior; speakers themselves are exposed to a wide range of partial samples of the full lexicon; and the experimental literature typically has not reported results at the fine granularity needed to evaluate certain mechanistic hypotheses (e.g. probability-matching) connecting speaker behavior to lexical structure. These issues have been at least partly addressed by *artificial language learning*, an alternative experimental approach reviewed in the following section which has informed the analysis of behavioral wug test data in this dissertation.

2.2.2 Regularization and speaker generalization in artificial languages

In artificial language learning (ALL) experiments, participants are taught a miniature language over the course of one or more training sessions, and then evaluated using test items. Many ALL experiments have focused on language acquisition (see Culbertson and Schuler, 2019, for a review) and the maturational constraints which cause language learning to differ for children and adults (e.g. Newport, 2016, 2020). This dissertation considers only adult behavior, where ALL appears comparable to second language learning (Ettlinger et al., 2016). The key advantage for ALL studies, however, is that the distribution of lexical and grammatical variants in an artificial language — as well as participants' exposure to those distributions — is under full control of the researchers. This means that linguistic phenomena of interest can be isolated

Singular Article + Noun	Plural
le dug	dugs
le gat	gats
le brid	brids
ze pik	piks
ze cheep	cheep
ze bish	bish

Test Input	A	B	C	D
le gee	gees	gees	gees	gee
ze koo	koos	koos	koo	koo
ze teer	teers	teer	teer	teers

Table 2.6: Hypothetical artificial language. Left table shows training vocabulary, right shows test outputs from four speakers. A regularizes the majority variant, B probability-matches, C regularizes conditional variation, and D regularizes the minority variant.

and manipulated with precision, and the behavior of individual speakers can be analyzed and classified in more detail (c.f. Chapter 6).

In a series of ALL experiments, Hudson Kam and Newport (2005) found that participants had a range of responses to morphological variation during training. After exposure to training sentences where an article would unpredictably appear with nouns in 60% of cases, some participants (mainly adults) *probability-matched* by reproducing this variation in novel test sentences, while other participants (mainly children) would either consistently produce or omit the article. Hudson Kam and Newport termed the latter behavior REGULARIZATION: by imposing consistency on an inconsistent grammatical system, these participants were effectively producing regularity, i.e. predictable lexical structure. Subsequent ALL research has developed more precise quantitative ways to define for regularization and probability matching, and characterize the experimental settings in which these behaviors occur.

Regularization is defined as a behavioral process which produces linguistic regularity. As such, it inherits some of the definitional vagueness discussed in §2.1.1, and the conceptual axes shown in Table 2.1 can clarify different approaches to defining regularization. As found by Hudson Kam and Newport (2005), speakers may regularize by producing (or omitting) one linguistic category consistently. In this understanding, regularization is a process which targets the distribution of one linguistic class considered in isolation. Others, however, define regularization as lowering the entropy of the overall distribution over linguistic classes (e.g. Smith and Wonnacott, 2010; Ferdinand et al., 2019). Table 2.6 presents the training and test vocabulary for a hypothetical ALL experiment, which I will use to illustrate these concepts. In Table 2.6, Speaker A regularizes in both the isolated and distributional senses; they observe the *-s* plural class on 60% of the training vocabulary, and apply this class consistently to 100% of the test forms. Probability-matching, by contrast, is inherently characterized in distributional terms. In Table 2.6, Speaker B probability-matches by variably producing the *-s* plural class for 60% of the test forms. Note that both regularization and probability-matching are defined in terms of distributions over output inflected forms here. There is, however, a third possibility: speakers may reduce unpredictable variation by *conditioning* on some other linguistic factor — in keeping with views of regularity which focus on input conditions. In Table 2.6, Speaker C illustrates this type of regularization: the singular article on the input noun (*le* or *ze*) is 100% predictive of their plural class assignments, although this relationship is less consistent in the training corpus. Speaker D highlights a potential source of conflict between the two definitions of regularization given here, to be clarified in later discussion: they are *regularizing with a minority variant* in the isolated sense, but *probability-matching* in the distributional sense. This section will review literature relevant to these behaviors, and consider how they relate to regularity and morphological generalization in natural language.

Regularizing with a majority variant One way that speakers can regularize is by increasing the frequency of one specific class in their productions. Typically, the linguistic class which is *most frequent* in the training corpus is targeted for relative overproduction on test items. Hudson Kam and Newport (2009) found that adult speakers overproduced the most frequent class when presented with variation between several articles (as opposed to variable production of one article, as in Hudson Kam and Newport, 2005); furthermore, the availability of more options (e.g. three determiner categories rather than two) led speakers to regularize even more. Note that this form of regularization is compatible with the distributional approach characterizing regularization as entropy reduction: in most cases, increasing the frequency of a majority variant corresponds to decreasing the entropy of the distribution. For example, in the hypothetical ALL experiment shown in Table 2.6, Speaker A increases the frequency of the majority class *-s* from 60% in the training distribution to 100% in the test distribution, reducing the entropy from .97 bits in training to 0 bits in test.

Focusing on how speakers generalize one particular class is consistent with some views of morphological regularity reviewed in §2.1, specifically the **rule generation** perspective, which considers linguistic properties of categories in isolation.[10] There is, however, a crucial difference

---

[10] The concatenativity view of regularity also considers categories in isolation; however, morphological variants in ALL experiments are typically controlled to be equal on this dimension, so concatenativity is not as relevant to

between these two literatures. In a rule-governed analysis of morphological regularity, for instance as developed in the generative tradition (c.f. §2.1.3), a particular linguistic category can generalize to novel forms only when licensed by the underlying linguistic system (i.e. if the novel form meets certain input conditions). If a speaker generalizes that linguistic category, this behavior reflects their implicit knowledge of their language's lexical structure. In the case of artificial language learning, however, speakers have no implicit knowledge of the specific language in question. For this reason, regularization in ALL experiments can only be attributed to general cognitive principles —- either domain-general biases, or biases specific to language but independent of any specific language. For instance, Hudson Kam and Newport's finding — that speakers regularize more in the presence of more variants — echoes earlier findings in non-linguistic domains (Gardner, 1957; Weir, 1964), suggesting that regularization may reflect domain-general cognitive load pressures. Reali and Griffiths (2009) and Ferdinand et al. (2019) find that increased cognitive load drives adult speakers to regularize more in both linguistic and non-linguistic tasks; however, speakers also regularize more in linguistic tasks overall. Culbertson and Kirby (2016) therefore propose that regularization reflects a domain-general cognitive bias toward *simplicity*, but the *strength* of the bias is domain-specific, i.e. stronger for language.

Regularizing with a minority variant In most cases, participants in ALL experiments regularize by overproducing the majority variant; however, many studies have also reported cases where one or more participants regularize by overproducing a minority variant instead (e.g. Hudson Kam and Newport, 2009; Reali and Griffiths, 2009; Smith and Wonnacott, 2010; Perfors, 2012, 2016). This is somewhat unexpected given the preceding discussion. In the case of natural language, one might expect speakers to generalize a minority linguistic class based on their knowledge of the underlying lexical structure, but ALL participants' knowledge of linguistic structure is limited to the sequences observed during training; it is unclear what kind of language-specific information would support this behavior. Ferdinand et al. (2019) identify one possible cause in their experimental data. They find that participants are more likely to regularize with a minority form when they observe it earlier during the training phase, suggesting that perhaps another domain-general cognitive bias — in this case, a primacy learning bias

---

this literature.

favoring early exposure — supports regularization with a minority variant.

Minority-variant regularization introduces an additional complication when regularization is quantified as entropy reduction. We can see this in the case of Speaker D, in the hypothetical experiment in Table 2.6. This speaker overproduces the minority class ø, assigning it to 60% of forms in test relative to 40% in training. The result, however, is a test distribution with .97 bits of entropy — exactly the same as the training distribution. If we look at entropy alone, Speaker D would be classified as probability-matching; however, the specific distribution of their test productions is also consistent with minority-variant regularization. Nuanced cases such as this illustrate the benefit of multiple perspectives on regularization, considering both individual class overproduction and entropy reduction (c.f. Ferdinand et al., 2019).

Regularizing by conditioning multiple classes So far, we have considered regularization behavior which takes the form of overproduction of one linguistic category relative to the training distribution, either the majority variant (exemplified by Speaker A in Table 2.6) or, less commonly, the minority variant (i.e. Speaker D). The third way to impose consistency on an inconsistent linguistic distribution, i.e. regularize, is to *condition* the distribution of multiple variants on particular linguistic contexts, as shown by Speaker C. We can quantify this type of regularization in terms of conditional entropy (Eq. 2.4). In the training corpus shown in Table 2.6, the entropy of the distribution over plural classes (*-s* or ø) conditional on knowing the noun's singular article (*le* or *ze*) is .46 bits. In Speaker C's test productions, however, that same measure is reduced to 0 bits, as C treats each test noun's singular article as fully predictive of its plural class. Note that this form of regularization is especially compatible with views of regularity as **predictability**. The Low Conditional Entropy Conjecture in particular (§2.1.2; Ackerman and Malouf, 2013) strongly implies that paradigmatic morphological organization arises due to speakers regularizing in this way (although this pattern may be driven by child rather than adult speakers).[11]

So do adult speakers regularize conditional variation? The ALL literature presents a mixed picture. Researchers have found that adults can successfully *learn* complex patterns of lexically-conditioned variation, i.e. cases where syntactic or morphological variants are used only with particular words (Wonnacott et al., 2008; Hudson Kam and Newport, 2009; Austin, 2010).

---

[11] Note, however, that most of the studies reported here also fail to find evidence for spontaneous conditional regularization by children (e.g. Ferman and Karni, 2010; Hudson Kam, 2015; Brown et al., 2022).

Smith and Wonnacott (2010) found that speakers introduced lexical conditioning of a variable plural marker — however, this conditioning developed over several generations of an *iterated learning* experiment, which will be discussed at greater length below. Adult ALL participants can also learn to condition on more general factors, from linguistic contexts such as syntactic role (Hudson Kam, 2015) or noun animacy (Ferman and Karni, 2010; Brown et al., 2022) to social contexts such as speaker identity (Samara et al., 2017) to purely distributional cues (Reeder et al., 2017). In a study of semantic conditioning, Brown et al. (2022) found that a few participants conditionally regularized — despite the partial consistency of this cue in training, they consistently applied semantic conditioning to novel nouns. This behavior, however, was atypical, occurring only when semantic cues during training were sufficiently early and salient to attract speakers' conscious attention. Across most studies, speakers typically condition by *matching* the conditional probabilities of the training data; they do not spontaneously exhibit the *additional* conditioning shown by Speaker C above.

A couple ALL experiments have used complex paradigm structures to assess the LCEC more directly. Seyfarth et al. (2014) trained participants on a small artificial language with three morphological suffixes marking number (singular, dual, and plural), and two noun inflection classes indicated by different number markers. They found that speakers' generalizations to novel words were influenced by both implicative relations between paradigm cells (I-complexity) and the type frequency of particular suffixes (E-complexity). Specifically, speakers used implicative paradigm relations to generalize suffixes with low type frequency, but did not appear to use this conditional information for suffixes with high type frequency. In a follow-up study, Johnson et al. (2020) tested paradigm learning rather than generalization, comparing the relative speed with which adult speakers and artificial neural networks (ANNs) learned paradigms with independently varying measures for I- and E-complexity. They found that, while the ANNs benefited from both lower I-complexity and lower E-complexity, speakers learned lower E-complexity paradigms more quickly but did not experience a similar advantage from lower I-complexity. Johnson et al. argue that implicative paradigm relations do not appear to be the key determinant of morphological complexity for human learners. Together with the studies discussed above, the overall evidence suggests that speakers are capable of learning and using conditional linguistic patterns under most circumstances, but rarely if ever spontaneously regularize conditional variation.

Probability-matching While speakers could impose consistency on a variable linguistic distribution using any of the three approaches to regularization discussed above, they can also simply reproduce the variability by *probability-matching* (c.f. Speaker B in Table 2.6). A substantial body of ALL research has found that, though some adult speakers regularize under some conditions, probability-matching is by far the most common behavior (see review by Newport, 2020). Like regularization, probability-matching is widely attested in non-linguistic domains; unlike regularization, probability-matching is difficult to motivate with appeal to broad cognitive principles. Probability-matching is fundamentally irrational behavior. It can never be the optimal strategy by definition (Icard, 2021), and it appears to demand more cognitive effort than regularization (Saldana et al., 2022). Nevertheless, while baboons, rats, and human children tend to regularize, pigeons, cockroaches, and human adults probability-match across a wide range of tasks and domains (Saldana et al., 2022).

Probability-matching is most closely associated with the **type frequency** view of morphological regularity, which predicts that speakers generalize inflection classes in proportion to their lexical frequency. In ALL experiments, probability-matching is often operationalized at the level of the individual speaker (e.g. Hudson Kam and Newport, 2009; Ferdinand et al., 2019), providing an analytical framework which could be usefully applied to studies of morphological generalization in natural language (c.f. Pierrehumbert, 2022). The operationalization, however, is not entirely clear in the case of *conditional* probabilities. As discussed above, various ALL studies have found that speakers do not conditionally regularize; however, they do learn and reproduce conditional probabilities. This raises questions for the application of probability-matching in a natural language setting, where a lexicon may contain many different potential conditioning factors. Does probability-matching in this case mean generalizing inflection classes according to their overall type frequency, or instead matching their lexical frequencies conditional on certain linguistic properties? I investigate this question in the context of German plural inflection in Chapters 6 and 7.

Communicative pressure vs. learner bias So far, we have considered how individual speakers might spontaneously regularize or probability-match relative to a fixed training corpus developed by experimenters. This type of evidence can inform our understanding of *learner biases*, meaning the cognitive tendencies that shape how individual speakers respond to particular experimental contexts. Natural language, however, reflects not only individual learner biases, but population-level patterns of cultural transmission. For example, consider a new word entering the lexicon: it must be encountered, learned, and used by a growing number of speakers in order to qualify as part of the language's vocabulary. Cultural transmission may exert its own *communicative pressures* independently from the biases of individual learners.

A number of ALL experiments have found evidence for these kinds of communicative pressures. One source of evidence comes from *interaction*. Dyadic interaction between participants in ALL experiments has been found to facilitate regularization in syntax (Fehér et al., 2016) and morphology (Rácz et al., 2020). Another source of evidence comes from *iterated learning* (Kirby, 2001; Kirby et al., 2014). In an iterated learning experiment, outputs from previous participants become the training corpus for new study participants, such that the training corpus evolves through interaction with several "generations" of speakers. Smith and Wonnacott (2010) found that, though participants probability-matched a variable plural marker in their training, after about five rounds of iterated learning this variation had stabilized and become *lexically conditioned*, such that participants produced each plural marker only with a specific noun. This finding suggests that conditional regularization might arise due to communicative pressures rather than biases at the level of individual learners.

Communicative pressures in transmission, then, provide an alternative hypothesis for the behavioral mechanism driving conditional patterns of morphological organization, as postulated by the Low Conditional Entropy Conjecture. The evidence for this mechanism, however, is far from conclusive. In the iterated learning domain, Smith et al. (2017) find that individual learner biases and communicative pressures can interact unpredictably. Cultural transmission can amplify weak learner biases in some cases, but it can also mask strong learner biases; moreover, the evolution of particular linguistic systems can be highly sensitive to non-linguistic aspects of the social context in which transmission occurs. Similarly, interaction does not reliably lead to conditional regularization. Fehér et al. (n.d.) predicted that dyadic interaction would facilitate semantic conditioning (based on noun animacy) for a variable plural marker. Instead, participants in their ALL experiment displayed either consistent lexical conditioning, or probability-matching; there was no evidence of conditional regularization for the higher-order semantic category. Communicative pressure remains a possible mechanistic hypothesis for the Low Conditional Entropy Conjecture, but we do not currently have evidence that cultural transmission reliably facilitates conditional regularization.

Summary In §2.2.2, I have reviewed findings from the experimental literature on *regularization* in artificial language learning (ALL). Speakers have various options for introducing regularity into variable linguistic distributions. They can regularize with a majority or minority variant, or regularize by conditioning multiple classes on particular linguistic contexts — although the latter behavior does not seem characteristic of individuals' spontaneous responses, but rather emerges through cultural transmission in some circumstances. Speakers can also probability-match, either to overall or conditional type frequencies. Researchers using the regularization framework have developed methods for fine-grained analysis of speaker behavior, including information-theoretic approaches to quantifying regularization and probability-matching at an individual level. Although natural language settings introduce additional complexity, ALL research methods can contribute useful insights to the study of morphological generalization.

In this dissertation, I apply the framework described above to analyze how German speakers generalize plural inflection. §3.3.1 reviews how these behavioral outcomes relate to the predictions of different linguistic theories, and Chapters 6 and 7 evaluate regularization and probability-matching behavior at the individual and aggregate level respectively.

2.3 Computational models of morphological generalization

By now, we've reviewed linguistic theories of morphological regularity, i.e. lexical structure (§2.1), and psycholinguistic studies connecting lexical structure to speaker behavior in natural and artificial languages (§2.2). Johnson (2017) relates these two domains to Marr's levels of description for cognitive systems (1982):

> The algorithmic level describes a cognitive system in terms of the representations and data structures involved and the algorithms that manipulate these representations. The computational level is the most abstract: it describes the goal(s) of the system, the information that it manipulates and the constraints it must satisfy. Linguistic theories are computational-level theories of language, while psycholinguistic theories of comprehension or production are algorithmic-level descriptions of how knowledge of language can be put to use. (Johnson, 2017, 172)

Despite this formal separability in principle, we have seen that many researchers consider these two domains to be fundamentally intertwined. This presumed connection has in turn informed computational modeling approaches to language. Johnson goes on to note that some statistical models of language rely upon the assumption that "the algorithmic level is derived from the computational level by general principles" (2017, 173; see also Smolensky and Legendre 2006 on isomorphism across levels of representation). This implies that a computational model which accurately represents linguistic structure should also at least partly capture language use, that is, speaker behavior.

In this section, I review proposed computational models of morphological generalization. These models all approach the task as conceived by Rumelhart and McClelland (1986): to learn the appropriate mapping from an input lexeme to an output inflected form, given a lexicon of natural language input-output pairs as training data. (Note that this task framing has been criticized as unrealistic from the cognitive standpoint of acquisition; see Ramscar, 2021, and discussion in §5.4. I nonetheless focus on this setup as the most direct computational equivalent of the behavioral wug test.) As these are considered models of linguistic cognition, the target model is one which shows human-like morphological generalization to unknown words; therefore, I focus on studies which evaluate model predictions through comparison to speaker behavior, rather than accuracy on a held-out test set. Though many models incorporate a range of influences, here I broadly group them according to the three morphological regularity criteria under consideration: rule generation, type frequency, and predictability (c.f. Table 2.1).

2.3.1 Symbolic models

Under the **rule generation** view of morphological regularity, regular morphological processes are best characterized in terms of rule application, where a rule specifies the structural transformation associated with a particular set of input conditions (c.f. Table 2.5). The class of computational models most closely associated with this view are *symbolic learners*. Many computational models include explicit symbolic structure of some kind; here, I mean specifically models which, given a lexicon, infer and apply these types of symbolic rules, i.e. mappings from input conditions to morphological transformations. We will consider two symbolic learners which have been applied to behavioral data on morphological generalization: the Minimal Generalization Learner (MGL) and the Abduction of Tolerable Productivity (ATP).

MGL The Minimal Generalization Learner (Albright and Hayes, 2003) is influenced by rule-based linguistic analyses as discussed in §2.1.3, but it differs in certain critical respects. The MGL *infers* rules of the requisite type; however, its *application* of those rules is probabilistic and parallel, rather than deterministic and serial (i.e. strictly ordered). The MGL posits rules at incrementally increasing levels of generality, starting with pairs of words and gradually building more general rules based on shared phonological and morphological features. Each rule is weighted by the frequency-adjusted scope of the input (i.e. how generally the rule applies, measured in the number of lexical forms which match the specified input condition) and the rule's *reliability*, i.e. its accuracy when applied to the selected forms. This enables the model to represent morphological processes with varying input scopes, including "irregular rules" which produce "islands of reliability" covering only small sets of words. To generate predictions, the MGL applies all rules simultaneously, with outcomes stochastically predicted according to the normalized sum over weights of applicable rules. Albright and Hayes (2003) found that adult speaker productions of English past tense forms in a wug test showed high correlations to MGL predictions.

The MGL crucially deviates from the rule generation view of morphological regularity in its treatment of input conditions. As discussed in §2.1.3, the generative view considers rule scope purely in terms of the abstract feature space of the input. For example, in Table 2.5, the input feature specification of rule R1 is a strict subset of the input feature specification of rule R2. Rules are prioritized strictly on the basis of generality in the input feature space; the number of observed instances which match R1 or R2 has no bearing (although presumably at least one instance in each feature combination is needed, so that the linguist can posit the rule in the first place). By contrast, the MGL counts the number of lexical instances which match the input conditions specified for a given rule, thus incorporating type frequency into its rule weighting. Veríssimo and Clahsen (2014) criticize the MGL on this basis, and propose an alternative Default Generalization Learner (DGL). This model extends the MGL by modifying its rule-weighting criterion, such that the rule with the least restrictions in the input feature space — i.e. the default rule — is automatically assigned the highest weight.[12] They analyze

---

[12] Veríssimo and Clahsen (2014) argue that this change renders the DGL a *dual-mechanism* model compared to the single-mechanism MGL, because the DGL draws a principled distinction between context-free (default) rules and context-sensitive (similarity-driven) rules, and does so independently of any statistical considerations. We will consider this claim in the broader context of dual-mechanism analyses in §3.2.1.

behavioral results from a verb inflection wug test with adult Portuguese speakers, and find a closer fit for DGL compared to MGL predictions.

ATP The Abduction of Tolerable Productivity (Belth et al., 2021) is a computational model based upon the Tolerance Principle (Yang, 2016, c.f. §3.2.1). The ATP recursively proposes and selects rules to yield a decision tree; see more detailed description in §4.3.1. The ATP's treatment of rule input scopes takes a somewhat intermediate approach between the MGL and generative analyses (e.g. §2.1.3). Rule *application* in the ATP is deterministic and serial, in contrast to the MGL. However, both the Tolerance Principle and the ATP rely upon the statistical criterion of accuracy during rule *inference* and *ordering*. In practice, an accuracy criterion is easily influenced by type frequency: the ATP will prioritize a rule that applies to $\frac{91}{100}$ words over a rule that applies to $\frac{9}{10}$ words, even though both rules readily pass the Tolerance Principle threshold. This statistical influence distinguishes the ATP approach from generative analyses which order rules based on input feature space alone. Belth et al. (2021) report a high correlation between the ATP predictions and adult wug test data from English (Albright and Hayes, 2003) and German (McCurdy, Goldwater and Lopez, 2020), noting crucially that their model can learn a default rule in the latter case. In Chapters 4 and 7, I further evaluate the fit between ATP predictions and speaker behavior, and investigate in particular whether default rule learning contributes to speaker-like German plural generalization.

The two symbolic learners considered here model morphological generalization as rule-based inference informed by a given lexicon. In this sense, both models accord with the view of morphological regularity as rule generation, including the emphasis on input conditions which characterizes that perspective. However, both models also incorporate elements from other perspectives on morphological regularity, especially predictability (in the form of accuracy), and, to a lesser extent, type frequency.

2.3.2 Frequency-matching models

As we saw in the previous section, **type frequency** is relevant to many modeling approaches, including symbolic rule learners; indeed, it's difficult to imagine any statistical modeling approach which could work *without* using lexical frequencies. Nevertheless, models vary substantially in the degree to which type frequency directly influences morphological generalization. For instance, the ATP (Belth et al., 2021) incorporates frequency only indirectly via the accuracy criterion. Artificial neural networks are another interesting case. While some linguists have criticized neural models for overreliance on type frequency in generalization (e.g. Marcus et al., 1995), Bybee criticizes them for *insufficient* reliance on the type frequency of output inflected forms, noting that "connectionist models do not form product-oriented generalisations" (1995, 433). We will return to artificial neural networks in §2.3.3. This section reviews two computational approaches which I consider more directly oriented to frequency-matching a given distribution, namely Bayesian models and frequency-weighted MaxEnt grammars.

Bayesian models Pierrehumbert (2022) observes that Bayesian models are particularly suited to modeling probability-matching behavior, as the posterior tends to converge on the data distribution. One illustration of this point is Perfors et al. (2010)'s successful application of a domain-general hierarchical Bayesian learner (Kemp et al., 2007) to model probability-matching behavior by adult speakers in an artificial language learning study (Wonnacott et al., 2008). Perfors et al. specified a model which could track statistical distributions at several hierarchical levels of analysis, such that it could form "overhypotheses" about feature variability. Separate instances of this model were trained on two different artificial languages: a *lexicalist* language in which each word appeared consistently with one of two noun particles, indicating stable membership in one inflection class, and a *generalist* language in which words appeared flexibly with both particles. When generalizing to novel nouns, the model showed frequency-matching behavior at multiple levels, similar to the human study participants. In particular, when exposed to the lexicalist language, both model and participants showed generalization patterns which were *lexically conditioned* at the level of individual words, but also *frequency-matched* at the level of the overall distribution. This result suggests that Bayesian models are especially well-suited to capturing probability-matching behavior.

The computational demands of fitting Bayesian models, however, makes it challenging to directly train them on categorical sequences, such as the strings of characters comprising words in a lexicon. This leads to difficulty applying Bayesian models to the problem of morphological generalization as we have framed it here, namely learning to map input lexemes to output inflected forms. One way to address this issue is to use simplified representations of the relevant data. For instance, Perfors et al. (2010) train their model on count data representing the number of times a word is observed with a particular class in the artificial language. Another approach is to include specific structure to handle symbolic sequences, typically rule-based generative grammars. For instance, Johnson et al. (2007) propose Adaptor Grammars, a class of nonparametric Bayesian language models which learn a distribution over the generation and retrieval of tree-structured strings generated by probabilistic context-free grammars (PCFGs); see also Fragment Grammars (O'Donnell et al., 2009), which extend the retrieval mechanism to cover subtrees. This approach, though powerful, can add significantly to the complexity of fitting a model, as the modeler must specify and explore the space of possible grammars. A third approach to Bayesian sequence modeling is to apply variational inference methods with neural networks (e.g. Graves, 2011; Chien, 2019), which removes the need for the modeler to explore the space, but can substantially increase the computational complexity of training the model. In the interest of expedience, I conduct preliminary Bayesian modeling experiments in this dissertation using the first approach of simplified data representations in §7.1.1.

Constraint-based models Linguists have also modeled probability-matching using weighted MaxEnt grammars. Maximum Entropy, or MaxEnt, grammars (Goldwater and Johnson, 2003) use the general statistical framework of additive log-linear modeling to weight constraints in Optimality Theory (Prince and Smolensky, 2004). Zymet (2019) uses a hierarchical regression framework to extend a MaxEnt grammar of variation in a constructed toy dataset. In this approach, lower-level random effects capture individual lexical idiosyncrasy while higher-level fixed effects capture overarching frequency distributions, leading to a model of hierarchically nested generalizations. O'Hara (2020) finds that an online MaxEnt learner with a lexical sampling scheme tends toward frequency matching as the size of the lexicon grows. As with some Bayesian models discussed above, this class of models requires the researcher to explicitly specify the grammar in question, here in the form of constraints. Due to my limited familiarity with the relevant constraint specifications, I do not use this class of models in the current work, but note their contribution to the literature on computational frequency-matching.

Optimality Theory, and by extension MaxEnt and related Harmonic grammars, stand in an interesting relation to the conceptual dimensions of morphological regularity described in Table 2.1. On the isolation-distributional axis, inflection classes are not evaluated in isolation, but only in connection with specific lexemes; the chosen output form is the one with the least constraint violation or most Harmony, which depends upon the distribution of generated forms. In this sense, evaluating morphological regularity is an inherently distributional process. On the input-output axis, Optimality Theory grammars explicitly represent this dimension as two competing families of constraints: Faithfulness (to the input) and Well-formedness (of the output). The task of the learner is to find a constraint ranking which maximizes the likelihood of the observed data. The presence of both input– and output-focused constraints in MaxEnt grammars might orient them more toward a view of regularity as predictability, rather than type frequency. However, the findings reported above suggest that, like Bayesian models, this class of models may tend to converge to the distribution of output forms observed in the data.

2.3.3 Exemplar-based and neural models

Finally, we consider computational approaches aligned with the **predictability** criterion of morphological regularity. Of course, accurate prediction, i.e. speaker-like generalization, is a goal for all of the models discussed so far. The distinguishing characteristic of models in this category is that they are *structure-agnostic*: while models reviewed in previous sections include explicit structure or mechanisms supporting a focus on input conditions (e.g. symbolic rule learners) or output type frequencies (e.g. Bayesian learners), exemplar models and artificial neural networks (ANNs) can freely and flexibly integrate any type of information supporting prediction. Of course, modeling design choices can strongly affect *how* different information sources inform generalization in neural or exemplar models; however, in contrast to models with more explicit structure, the opaque inner workings of neural models in particular mean that these design choices do not transparently correspond to categories of linguistic interest (Baroni, 2021). Throughout this dissertation, unless otherwise specified, I will use "neural model" to mean deep neural models with one or more hidden layers.

The opacity of ANNs raises many additional questions, including the nature of the relationship between exemplar and neural models, both of which are said to use *similarity* or *analogy* in generalization. Are these two model classes really doing the same thing? Or do they differ, and if so, how? Hahn and Chater (1998) compare similarity– and rule-based generalization, and conclude that, strictly speaking, neural models do neither: they do not explicitly represent abstractions, as rule-based models do, but they also do not explicitly represent data instances, which are required to compute similarity under exemplar models. Ashby and Rosedahl (2017) show mathematical equivalence between an exemplar model (the GCM, discussed below) and one particular biologically-plausible neural model, with unclear implications for the very different ANN architectures used in practice. Ambridge (2020) argues that neural models fall on a spectrum between abstract and exemplar-based computation, partly as a function of capacity; given a sufficiently large number of hidden units, a deep neural model can become a 'de facto exemplar-based model' by individually representing each instance. He claims that many recent computational successes of highly parameterized neural models are attributable to this exemplar-driven computation, although Mahowald et al. (2020) counter by highlighting these models' abstract representational capacities. Dasgupta et al. (2021) present a behavioral method to distinguish rule– and exemplar-based generalization. They find that modern ANNs largely interpolate between these two extremes, further supporting a spectrum of abstraction; however, they also report an overall bias toward exemplar-based generalization, especially for out-of-distribution stimuli. Neural models are clearly related, but not reducible, to exemplar models. Here, we consider how these approaches have been used to model morphological generalization, still focusing on studies which compare with speaker behavior.

Exemplar models Three prominent exemplar models which have been applied to inflectional morphology are the Generalized Context Model (GCM; Nosofsky, 1988), the Analogical Model (AM; Skousen, 1989), and the Tilburg Memory-Based Learner (TiMBL; Daelemans, 2002); see Chandler (2017) and Ambridge (2020) for review and detailed comparison. In the domain of English past tense inflection, Albright and Hayes (2003) find that the rule-based MGL (c.f. §2.3.1) better predicted speaker behavior than the GCM, though Chandler (2010) reports an even closer fit for AM predictions on the same data. Rácz et al. (2020) use the GCM and the MGL to model speakers' English past tense productions for novel words in an interactive task, and find independent effects from both, though the GCM better captures behavioral changes in speaker alignment over the course of the experiment. Keuleers et al. (2007) show that TiMBL can model how Dutch speakers generalize plurals by integrating multiple sources of information, i.e. both the phonological and orthographic form of the novel noun. This finding illustrates how exemplar-based models can flexibly incorporate any information to support prediction, in keeping with a predictability-based view of morphological regularity. In Ch. 7, I evaluate a simple nearest-neighbor exemplar model following Blevins et al. (2017, c.f. §2.2.1).

On the other hand, this flexibility can impede accurate cognitive modeling in cases where speakers *don't* use all of the available information. Dawdy-Hesterberg (2014) uses the GCM to model wug test data on nominal and verbal inflection from Arabic speakers. She finds that the GCM accurately predicts speaker behavior for nouns but not verbs. In the latter case, the GCM relies on fine-grained segmental data to make confident predictions for verb masdar patterns, but speakers appear to probability-match lexical statistics at the more coarse-grained level of the prosodic template. Similarly, Nieder et al. (2020) find that lexical statistics predict the plural generalizations of Maltese speakers, outperforming a *discriminative learning* model (Nieder et al., 2022)[13] — another model in the exemplar family, which we will now consider.

Discriminative learning models (e.g. Baayen, 2011; Ramscar et al., 2013; Baayen and Hendrix, 2017; Ramscar, 2021) are effectively shallow neural networks — that is, a neural model with an input and output layer, but no intervening hidden layer. As such, this model class represents an intermediary point between the purely memory-based systems considered above, and the deep neural models to be discussed below. Discriminative learners do not store each individual instance encountered during training; instead, they update the weighted connections betwen input and output layers using an error-driven learning rule. Nonetheless, Ambridge (2020) argues the cue-weighting learned by these models is strictly equivalent to the feature-weighting found in many exemplar models (though c.f. Skousen's AM). Ambridge further identifies the hidden layer of deep neural networks as necessary to represent abstract mappings (i.e. rule-like generalization); because discriminative learners lack this capacity, they must generalize through analogy to exemplars.

One crucial factor which distinguishes exemplar-based (including discriminative) models from deep neural network models is the need for explicit feature extraction in the former case. Exemplar models require relevant features to be directly available in order to compute similarity. Feature selection and extraction can be part of the data preparation by researchers. For instance, Keuleers et al. (2007) prepare prosodically segmented phonological and orthographic representations of the Dutch noun lexicon for their model, and Ramscar et al. (2013) extend a corpus of child-directed speech with semantic cue bundles (e.g. "mice" would be annotated

---

[13] Nieder et al. present their findings as supporting the cognitive modeling capabilities of the Naive Discriminative Learner (NDL, Baayen, 2011), but they also report that simple lexical frequency counts show a correlation of .86 to speaker productions, which appears to substantially improve on all tested NDL versions (reported correlations .65–.77) (2022, Table 3).

with labels like "mousiness, multiple-items, mouse-items") to support discriminative learning. The feature extraction process can also be automated, as in the case of AM's supracontexts, which are generated by deleting sequence elements (e.g. the orthographic word "mice" would yield supracontexts "mic_", "mi_e", "m_ce", "_ice", "mi__", "m__e", "__ce", and so on). In any case, the researcher generally specifies which features are available for similarity computation in an exemplar-based model. In contrast, deep neural networks typically operate directly on sequential linguistic data. This effectively automates the process of feature extraction, making the modeling process simultaneously more convenient and less interpretable.

Neural network models In 1986, Rumelhart and McClelland initiated decades of debate with a simple research question: can a statistical model trained on a natural language lexicon learn the inflectional categories present in that lexicon, and furthermore, learn to generalize those categories as human speakers do? Theirs was one of the first works to apply deep Artificial Neural Networks (ANNs) to the task of modeling linguistic generalization. Over the past decade, researchers have achieved great progress in modeling natural language using ANNs, with some models arguably approaching human-like competence in formal linguistic domains such as syntax (Mahowald et al., 2023). In the domain of morphology, ANNs have decisively outperformed other model classes in SIGMORPHON inflection-modeling challenge tasks, which are evaluated by predicting held-out lexical forms (e.g. Cotterell et al., 2016, 2017; Cotterell, Kirov, Sylak-Glassman, Walther, Vylomova, McCarthy, Kann, Mielke, Nicolai, Silfverberg, Yarowsky, Eisner and Hulden, 2018). Kirov and Cotterell (2018) trained an encoder-decoder recurrent neural network on English verb inflection, and found that its predictions on novel verbs correlated closely to behavioral data collected by Albright and Hayes (2003). Based on this finding, along with the broader successes of ANNs in modeling inflection, Kirov and Cotterell propose that modern ANNs have resolved Rumelhart and McClelland's question in the affirmative, by demonstrating speaker-like morphological generalization. Both of these points, however, have been challenged. Careful evaluation indicates that neural models have not yet fully solved the task of generalizing inflectional morphology (Liu and Hulden, 2021; Goldman et al., 2021). Furthermore, Corkery et al. (2019) use the same modeling approach with various random seeds, and find that ANN predictions do not show reliably high correlations to Albright and Hayes's data. These developments have inspired more cognitively-focused challenge tasks with evaluation on behavioral data at recent SIGMORPHONS (e.g. Pimentel et al., 2021; Kodner and Khalifa, 2022).

The power and flexibility of neural models make them a natural fit with the regularity-as-predictability view of morphology. Consistent with the Low Conditional Entropy Conjecture (LCEC), ANNs have been used to implement and study implicative relations within morphological paradigms (Malouf, 2017). Cotterell, Kirov, Hulden and Eisner (2018) use ANNs to implement an extended version of the LCEC in which inflectional paradigms are generated as tree-structured Bayesian networks. Their typological survey supports Ackerman and Malouf (2013)'s postulated trade-off between E-complexity (enumerated number of unique morphological forms) and I-complexity (conditional probability). Wu et al. (2019) use ANNs to directly assess the predictability, or regularity, of inflected forms from held-out lexemes for 28 different languages, and find that low predictability (i.e. irregularity) correlates with token frequency across languages. The findings of these studies, which use ANNs to model the predictability of morphological inflection across typologically distinct languages, are consistent with long-standing observations from the linguistics literature, and indicate that ANNs are capable of capturing fine-grained conditional or implicative morphological relations. We cannot necessarily assume, however, that these conditional relations will have the same degree of influence on morphological generalization in speaker behavior. In a comparative artificial language learning study, Johnson et al. (2020) find that ANNs benefit from implicative relations within morphological paradigms — lower I-complexity speeds up learning — but the same benefit does not appear for human speakers.

In this dissertation, I use the recurrent encoder-decoder architecture (ED; Kann and Schütze, 2016) proposed as a cognitive model by Kirov and Cotterell (2018) to model German plural generalization. This represents only one point in the vast design space of neural network architectures, which display a considerable range of inductive biases relevant to generalization performance (e.g. Kharitonov and Chaabouni, 2020; McCoy et al., 2020). While it arguably limits the scope of claims emerging from this dissertation, the decision has two motivations. The first reflects what Baroni (2022) calls "low commitment to models:" model selection in deep learning research is typically guided by application performance rather than theoretical considerations, making it difficult to interpret the theoretical significance of variation in neural architectures. This issue is further compounded by the vast design space of possible architectures. To keep the scope of the research tractable, I consider only the ED, which Kirov and Cotterell (2018) motivate with in-depth comparison to the original model proposed by Rumelhart and McClelland (1986) for the English past tense. The second reason is a general intuition that the phenomena of interest in this thesis reflect properties shared by statistical learners broadly. For instance, the lexical relationship between for grammatical gender and plural class is sufficiently strong to be detected by nearly any statistical learner, as reflected in results from a wide variety of both neural (e.g. Goebel and Indefrey, 2000; Williams et al., 2020; Dankers et al., 2021; Beser, 2021) and non-neural architectures; see §3.4 for further discussion. For these reasons, I treat the ED as a stand-in for neural network models more generally in comparative analysis with other classes of statistical learners.

Summary This section has reviewed computational models of morphological generalization. I have focused on the inflection modeling task as framed by Rumelhart and McClelland (1986), to learn a mapping between input lexemes and output inflected forms, given a lexicon as training data. This restricted scope means I have excluded approaches which seek to model communicative pressures or cultural transmission, such as iterated learning or agent-based models (e.g. Kirby, 2001; Reali and Griffiths, 2009; Smith and Wonnacott, 2010; Kirby et al., 2014; Smith et al., 2017; Ferdinand et al., 2019; Round et al., 2022). We have considered computational approaches in terms of their affinity with linguistic perspectives on morphological regularity (§2.1): symbolic learners (§2.3.1) align with the rule generation account, Bayesian and constraint-based learners (§2.3.2) align with the type frequency account, and exemplar-based and neural models (§2.3.3) align with the predictability account. In Chapter 7, I compare models from each of these classes in terms of their fit to behavioral data.

2.4 Chapter summary

This chapter has reviewed linguistic (§2.1), psycholinguistic (§2.2), and computational (§2.3) approaches to characterizing morphological *regularity*, conceived in terms of mapping from input lexemes to output inflected forms. We have reviewed three broad perspectives on morphological regularity — in terms of input-focused rule generation, output-focused type frequency, and integrative predictability — which offer competing explanatory accounts of the lexical structure driving morphological generalization. We have also reviewed the behavioral literature on how speakers generalize inflection in experimental conditions with novel words (§2.2.1), and considered the potential contributions of the *regularization* framework developed in recent literature on artificial language learning (§2.2.2). Finally, we've considered how different computational approaches to modeling morphological generalization relate to perspectives on regularity, connecting rule generation with symbolic learners (§2.3.1), type frequency with Bayesian and MaxEnt learners (§2.3.2), and predictability with exemplar-based and neural network learners (§2.3.3). Chapter 3 reviews how these linguistic criteria for morphological regularity have informed theoretical, behavioral, and computational research on the German plural inflection system.

Chapter 3

Background: German Plural

As the preceding chapter has shown, a great deal of research on morphological generalization has been conducted within the context of the *past tense debate*, with many studies focused on English verbal inflection (e.g. Rumelhart and McClelland, 1986; Pinker and Prince, 1988; Prasada and Pinker, 1993; Pinker and Ullman, 2002; Albright and Hayes, 2003; Chandler, 2010; Seidenberg and Plaut, 2014; Blything et al., 2018; Kirov and Cotterell, 2018; Corkery et al., 2019; Rácz et al., 2020). Although there is certainly some complexity to this problem (as attested by the existence of all these studies), the English past tense is a relatively simple inflection system globally speaking, with one statistically predominant regular inflection class. This fact makes English verbal inflection particularly ill-suited to evaluate models of morphological regularity (c.f. Table 2.1): different criteria — i.e. default rule generation, type frequency, and predictability — all converge to predict the regular past tense suffix *-ed*. Models of English past tense production must be compared on the relative ordering of a handful of low-probability irregular forms, which yields unstable and inconclusive analyses (Corkery et al., 2019). This suggests that morphological regularity might be better evaluated in a linguistic environment with sufficient complexity to distinguish between competing theories.

Marcus et al. (1995) propose resolving this issue through evaluation on the famously complex German plural system, which for decades has resisted straightforward linguistic description (Bloomfield, 1933; Wurzel, 1970; Mugdan, 1977; Augst, 1979; Bartsch and Vennemann, 1983; Dressler et al., 1987; Janda, 1990; Köpcke, 1993). Crucially, German has no majority plural class. This means that the three criteria in question make *divergent* rather than convergent predictions for morphological regularity — leading to cacophonous debate in the research literature, as we shall see.

In this chapter, we begin with a basic overview of the German plural system and some relevant lexical statistics (§3.1). We then consider how the linguistic (§3.2), psycholinguistic (§3.3), and computational (§3.4) approaches to morphological regularity reviewed in Chapter 2 have been applied to German plural inflection, and finally, how this literature motivates the research questions explored in this dissertation (§3.5).

3.1 The German noun lexicon

Each noun in the German lexicon has two static lexical attributes: grammatical gender and plural class. Grammatical gender can be masculine, feminine, or neuter, and is expressed on the definite article preceding a singular noun, e.g. *der Hund* "the (masc.) dog", *die Kuh* "the (fem.) cow", *das Kind* "the (neut.) child". This gender distinction on the article is collapsed for plural inflection: all plural nouns take the definite article *die*, homologous with the singular feminine article. The plural class of a noun is expressed by the form of the plural inflected noun itself, and characterized by how it differs from the singular form of the noun.[1]

The German plural system comprises five main suffixes: -*e*, -*er*, ∅, -*en*, and -*s*. The first three can optionally combine with an umlaut over the root vowel. Umlaut is a process which fronts a back vowel, so only roots with back vowels can take an umlaut (e.g. *Dach* → *Dächer*, *Fuss* → *Füsse*). Umlaut varies semi-independently of plural suffix, and is not fully predictable. While the phenomenon of plural umlaut has been addressed in the linguistic (Wiese, 1996; Trommer, 2020) and modeling (Wulf, 2002) literature, for simplicity, this dissertation will focus only on suffixes, taking them to define five separate plural classes. Table 3.1 gives example forms for each plural suffix, along with its **type frequency** (counting each word type only once, how many types in the lexicon take this plural?) and **token frequency** (how often do words with this plural suffix appear in the corpus overall?) as recorded in the CELEX lexical resource (Baayen et al., 1995; Sonnenstuhl and Huth, 2002), which is the source of all lexical statistics reported here. Note that, while -*en* and -*e* are the most frequent suffixes, there is no majority plural class over all nouns.

---

[1] We here follow the literature in focusing on the nominative-case form of each noun, considered the citation form.

Suffix	Singular	Plural	Type	Token
-en	Strasse	Strassen	48%	45%
-e	Hund Kuh	Hunde Kühe	27%	21%
∅	Daumen Mutter	Daumen Mütter	17%	29%
-er	Kind Wald	Kinder Wälder	4%	3%
-s	Auto	Autos	4%	2%

Table 3.1: German plural system with examples, ordered by CELEX type frequency (Sonnenstuhl and Huth, 2002).

3.1.1 Lexical distributions

Various linguistic factors interact with the lexical distribution of German plural classes:

- **Lexical: Grammatical gender**. There is a strong association between grammatical gender and plural class; in particular, almost all feminine nouns take *-en*.
- **Phonology: Final syllable**. Certain word endings are highly predictive of plural class: most words ending in schwa take *-en*, most words ending in a reduced final syllable (-er, -el, -en) take zero, and most words ending in a full vowel (-a, -o) take *-s*.
- **Morphology: Derivation and Compounding**. Particular derivational suffixes consistently take a specific grammatical gender and plural inflection class. For instance, any noun with the derivational suffix *-ung* always has feminine gender and takes the *-en* plural class, e.g. *die Reservierung* — *Reservierungen* "reservation — reservations", *die Sendung* — *Sendungen* "transmission — transmissions." Similarly, compound nouns take the grammatical gender and plural class of the rightmost ("head") noun: *das Kind* — *Kinder* "the child — children", *das Kleinkind* — *Kleinkinder* "the small child — small children." This morphological inheritance relation is highly systematic and nearly exceptionless in the lexicon. Note that this morphological relation further reinforces the statistical associations with gender and phonology listed above, as derived forms always share the same gender, final syllable, and plural class.
- **Semantics**. Proper names take *-s* (Marcus et al., 1995), and *-e* or *-en* if *-s* is phonologically blocked (Indefrey, 1999). There are also semantic animacy patterns which appear

	$H(C)$	$MI(C;G)$
All, 6 classes	1.98	0.67
Williams et al. 2020, 16 classes	2.88	0.71

Table 3.2: Entropy measurements over 6 plural classes $H(C)$ and mutual information between plural class and gender $MI(C;G)$. Compare reported measurements from Williams et al. (2020, Table 3) over a 16-class grouping of the same corpus.

	$\|F,G\|$	$H(F,G)$	$H(F)$	$H(G)$
-e	47	4.57	3.58	1.14
-en	50	3.00	2.56	0.57
-er	24	3.15	2.79	0.55
ø	48	2.48	1.75	0.93
-s	53	4.94	3.76	1.39
other	35	3.55	2.88	1.51
All	71	4.67	3.53	1.52

Table 3.3: Per-plural-class and overall entropy measurements for final character (F) and grammatical gender G. $|F,G|$: number of observed feature combinations. Entropy calculations are based on the observed frequency of feature values within the lexemes in a particular plural class, e.g. in the top row $H(F,G) = H(F,G \mid \textit{-e}), H(F) = H(F \mid \textit{-e})$, etc.

to interact with grammatical gender (Gaeta, 2008). Semantic factors are out of scope for this dissertation, but recent behavioral evidence indicates they may only weakly influence speaker plural generalization (Heitmeier and Frank, 2021).

Figs. 3.1 and 3.2 visualize the conditional distribution of plural classes in terms of the first two listed factors, namely grammatical gender and the phonology of the final syllable (represented in simplified form here as the final character of the orthographic singular form of the noun). As these plots show, these two features are highly predictive of plural class overall, although some feature combinations are more predictive than others. Another way to consider this is through a simple majority-class baseline: if we were to simply assign each noun in CELEX the majority plural class observed for that grammatical gender and final character combination, this would predict the correct plural class for 81.5% of nouns in the lexicon.

As discussed in §2.1, different criteria for morphological regularity are associated with different distributional properties; for instance, the default rule generation criterion connects regularity with maximal variability in the input feature space (§2.1.3), while the predictability criterion instead focuses on minimizing conditional entropy (§2.1.2). Here, I provide a brief

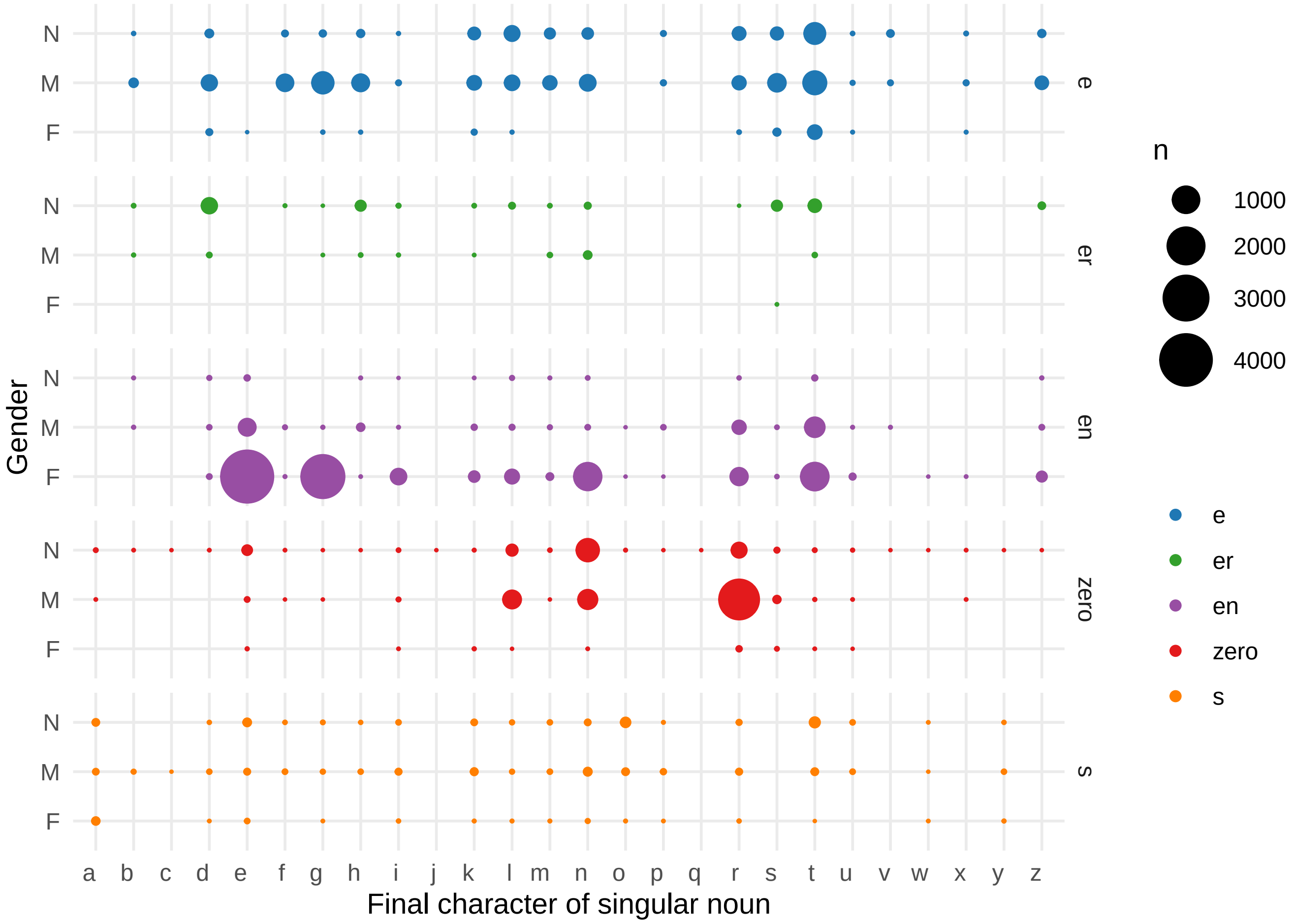


Figure 3.1: Distribution of plural classes for 25,999 German nouns in CELEX2 (Baayen et al., 1995), grouped by grammatical gender and the final character of the singular noun. Bubble size indicates number of nouns. Though the *-s* plural class is rare overall, it appears with 53 unique character-gender combinations, a more diverse set of environments than any other plural class.

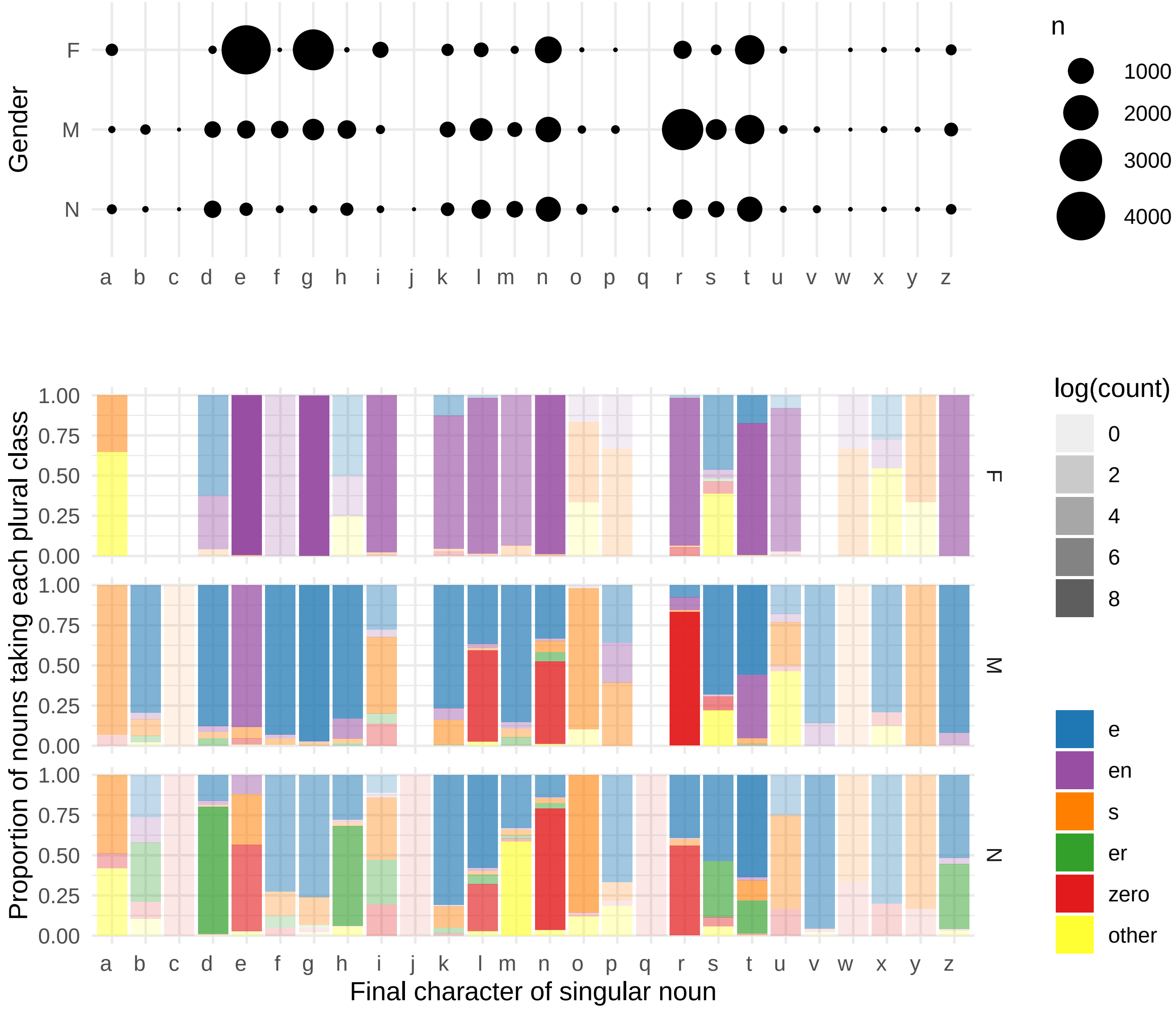


Figure 3.2: Distribution shown in Fig. 3.1, factorized into the joint distribution of gender and final character ($p(x)$, above) and plural class conditional on those contexts ($p(y|x)$, below). Higher observed counts for input feature combinations (upper plot) are represented as higher opacity (lower plot). While some feature combinations are uniformly associated with a particular plural class (e.g. feminine nouns ending in *-e)*, others show more variable distributions (e.g. nouns ending in *-t*, or nonfeminine nouns ending in *-n*).

information-theoretic description of the relevant lexical distributions so that these quantities can inform the theoretical discussion of German plural inflection in §3.2. Table 3.2 shows the information-theoretic relationship between grammatical gender and plural class over all nouns in CELEX2, using calculations based on type frequency. The overall entropy for plural class $H(C)$ and gender $H(G)$ is calculated with Eq. 2.1. The conditional entropy of plural class given gender $H(C \mid G)$ is calculated with Eq. 2.4. For conceptual ease, I report this relationship in terms of mutual information $MI(C;G)$:

$$MI(C;G) = MI(G;C) = H(C) - H(C \mid G) \tag{3.1}$$

As this equation shows, maximizing mutual information is definitionally equivalent to minimizing conditional entropy, so this quantity is relevant to the predictability account of regularity. For comparison, Table 3.2 includes the same values calculated by Williams et al. (2020). Note that their analysis is not directly comparable, as they use the 16 inflection class labels provided by CELEX, which mark out umlauts, loanwords, and various other distinctions; it is unsurprising that they find a higher entropy value $H(C)$ compared to an analysis based on 6 simplified classes. Despite this, they report a similarly high degree of mutual information with grammatical gender.

Entropy can also be used to compare the *contextual variability* of plural classes, or how broadly they are distributed across different linguistic contexts; as mentioned above, variability in the input feature space is relevant to the rule generation criterion for morphological regularity. Table 3.3 presents the number of unique feature combinations (in terms of grammatical gender G, and the final character of the singular noun, F) in which each plural class appears, along with the entropy measurements for each feature's distribution with respect to that plural class. We can see that *-er* appears in the lowest number of unique contexts $|F,G|$, but ø appears to have the most restricted distribution as measured by the joint entropy over contexts $H(F,G)$. On both of these measures, *-s* shows the least restricted distribution of all plural suffixes.

3.2 Linguistic theories of German plural inflection

A range of linguistic theories have been proposed to account for the plural distributions reviewed above. Here, I will review three accounts which featured prominently in past tense debate —

minority default theory (§3.2.1), gender-conditioned default theory (§3.2.2), and schema theory (§3.2.3) — and discuss how they relate to different morphological regularity criteria (Table 2.1) and the experimental work presented in later chapters.

3.2.1 Minority default

The complexity of the German plural system makes it possible to distinguish between competing accounts of morphological regularity (Table 2.1). This is not the case for English verbal inflection, where the past tense suffix *-ed* qualifies as regular both in terms of its *output* distribution (i.e. having the highest type frequency) and its *input* distribution (i.e. appearing in the most diverse set of linguistic contexts). In the German plural system, however, these two criteria can be dissociated. Consider the respective type frequencies (Table 3.1) and contextual variability measurements (Table 3.3) of the plural classes. *-en* has the highest type frequency, appearing with nearly half of the nouns in CELEX. Unsurprisingly, *-en* also appears in a wide range of input environments, with 50 unique gender and final character combinations. The more surprising observation is that the plural class *-s* appears with an even wider range of 53 feature combinations, despite its low type frequency of 4%. Furthermore, *-s* is the only plural class with a more variable distribution over linguistic contexts (measured by joint entropy: $H(F, G) = 4.94$) than seen in the lexicon as a whole ($H(F, G) = 4.67$). The variability of the *-s* plural extends beyond lexical contexts (Janda, 1990), as summarized by Clahsen:

> The -s plural applies when the phonological environment does not permit any other plural allomorph. It occurs on masculine, feminine, and neuter nouns, on words that exhibit the canonical stress pattern and on those that do not, on monosyllables and polysyllables, and on both vowel-final and consonant-final stems. The -s plural also generalizes to rootless and head-less nouns, for example, to nominalized conjunctions such as die Wenns und Abers, "the Ifs and Buts," to eponyms and product names (Fausts, Golfs, etc.), and to nominalized verb phrases (VPs) (die Rührmichnichtans, "the Touch-me-nots"). (Clahsen, 1999*b*, 995)

Based upon these observations, Marcus et al. (1995) argue that the *-s* plural class is *elsewhere-distributed* (c.f. §2.1.3) and therefore a MINORITY DEFAULT (in contrast to a default with a statistical majority in type frequency, such as the English past tense *-ed*). Under this analysis, -*s* is *productive* — i.e. generalizable to novel nouns — due to being *rule-generated*, specifically by an Elsewhere-Conditioned rule which permits any input noun to take -*s*. Marcus et al. (1995) additionally invoke the *concantenativity* criterion for morphological regularity. Of the five German plural classes, -*s* arguably imposes the least degree of phonological transformation between input and output: it is not associated with any segmental changes to the noun stem (unlike -*e*, -*er* and *ø*, which can combine with umlaut), and it does not change the number of syllables of the noun (unlike -*en*, which adds a syllable for any noun which does not end in a reduced final syllable such as -*er*, -*el*, or ə). From the convergence of these two criteria — rule generation and concatenativity — Marcus et al. (1995) conclude that -*s* is morphologically regular in the German plural system.

Note that both of the regularity criteria associated with -*s* are properties of one inflection class considered in isolation (c.f. Table 2.1). Are other German plural classes also regular? Logically, there must be mechanisms to generalize the other classes; if -*s* is the only productive class, we would expect the German plural system to look much more like the English plural system. Here, we consider two different theoretical accounts of German plural regularity which both take -*s* as the minority default. The Dual Mechanism Model (DMM; Marcus et al., 1995; Clahsen, 1999*b*) posits that -*s* is likely the only rule-generated plural class, and most or all other classes are generalized based on similarity. By contrast, the Tolerance Principle analysis (TP; Yang, 2016) posits that other plural classes are rule-generated as well. In any case, regardless of the specific proposals, the minority default analysis of -*s* continues to influence recent work in theoretical linguistics (e.g. Trommer, 2020; Schuhmann and Putnam, 2021).

Dual-Mechanism Model

The dual-mechanism, or dual-route, model draws a theoretical distinction between *storage* and *computation*: irregular wordforms are stored in the lexicon, while regular forms are computed by concatenative rule (Pinker and Prince, 1988). These two pathways are claimed to reflect the independent cognitive processes of declarative memory and procedural computation respectively (Pinker and Ullman, 2002). Of course, analogical comparison is also a form of computation, but it is one that relies on stored memory; rule-based computation is by definition memory-independent.

Marcus et al. (1995) argue that the elsewhere distribution of the *-s* plural suffix in German indicates that it is regular, i.e. generated by rule. In contrast, the other plural classes are productive on the basis of memory alone — that is, they can only generalize to novel words which are sufficiently similar to exemplars stored in the lexicon. This analogical generalization is necessarily constrained by existing memory, while the rule generating *-s* is fully general and memory-independent, with no input restrictions. Following the logic of the Pāṇini principle (§2.1.3), the more specific process of analogical generalization takes precedence over the more general Elsewhere-Conditioned default rule. Therefore, the memory-based generalization of other plural classes can preempt the rule-based generalization of *-s*, letting it remain a minority default. Marcus et al. (1995) allow that other plural classes may be morphologically regular, i.e. rule-generated, in particular contexts; for example, *-en* applies predictably to nouns which are feminine and end in a reduced final syllable (c.f. Fig. 3.2). These contexts, however, are necessarily constrained, and thus also preempt *-s* default rule application.

A crucial implication of this analysis is that any model which does not explicitly include rules, but generalizes based only on similarity to items stored in memory, should struggle to learn the correct pattern of *-s*-generalization in German:

> [P]attern associators neither easily generalize low-frequency suffixes, nor unite the different default circumstances (phonological and derivational) as defaults. [...] These models predict that -s should be eschewed across the board: driven only by phonological similarity, the models should always prefer the more common -e, -en, and -er plural forms [...] (Marcus et al., 1995, 233)

From this reasoning, Marcus et al. argue that artificial neural networks and other single-mechanism similarity-based models, i.e. "pattern associators," should fail to learn speaker-like generalization behavior in the domain of German plural inflection. They provide behavioral evidence to support their analysis (§3.3.2), but no computational evidence to support the claim posed above. I review the existing computational literature on this question in §3.4, and investigate Marcus et al.'s hypothesis directly with behavioral and computational experiments in Chapter 4.

Rule Number	Input Condition	Structural Change	Ex: Singular	Ex: Plural
1	[+fem]	*-en*	die Nation	die Nationen
2	[+RFS] (*-ər*, *-əl*, *-ən*)	*ø*	das Fenster	die Fenster
3	[+-ə]	*-(e)n*	das Auge	die Augen
4	[+monosyllabic, +neuter, +back vowel] (/a,o,u/)	*-er*	das Wort	die Wörter
5	[-fem] / [+masc/+neut]	*-e*	der Tisch	die Tische
6	[]	*-s*	das Auto	die Autos

Table 3.4: Productive rules of the German plural according to Yang (2016). RFS indicates Reduced Final Syllable, where the singular form ends in *-ər*, *-əl*, or *-ən*.

Tolerance Principle

Yang (2016, Chapter 4) analyzes German plural inflection using the methodology of the Tolerance Principle, described in §4.3.1, applied to an unreleased dataset of plural forms from German child-directed speech. Recursive application of the Tolerance Principle derives rules 1–5 in the order shown in Table 3.4. Curiously, rule 6 — the default rule which generates *-s* — is not derived through application of the Tolerance Principle (which requires a certain frequency threshold, as discussed above), but rather through appeal to the Elsewhere Condition: "unlike rules [1–5], which all refer to gender, the *-s* suffix places no restrictions on gender, and thus is most general, making it the default rule" (Yang, 2016, 134). This is unexpected, as earlier in the book, Yang claims that the minority default *-s* *can* be derived through the Tolerance Principle criterion for productivity:

> [T]he German plural system (section 4.4), can have a productive rule ("add *-s*") that applies to very few nouns because morphosyntactic and phonological features help partition nouns into several classes, all of which contain productive rules within, and thus do not constitute exceptions to the *-s* suffix. (Yang, 2016, 74)

Although Yang's own analysis does not actually substantiate this claim, Belth et al. (2021) demonstrate that it holds under some conditions. They extend the Tolerance Principle by adding a search mechanism to yield the ATP (c.f. §2.3.1, §4.3.1). They use the ATP to model an openly-released dataset of plural forms from German child-directed speech, and find that the model learns an *-s*-default rule in about 20% of their simulations, where each simulation samples a different subset of the data. This finding is reproducible, if limited in scope: the largest dataset they consider has 400 nouns. Chapter 4 investigates whether the ATP model still achieves -*s*-default analyses for datasets at larger though still human-vocabulary-level scale,[2] and whether this contributes to speaker-like generalization.

On dual mechanism models broadly

We conclude the minority default discussion by considering dual mechanism models as a broader category. While the DMM described above (e.g. Pinker and Ullman, 2002) is the most well-known proposal in this vein, other models under consideration here also make interesting use of dual mechanisms. For instance, Yang's TP analysis characterizes morphological generalization as realized exclusively through rule application, but in practice the ATP implementation falls back on analogical generalization from stored exceptions if it fails to learn a productive rule for a given context (Belth et al., 2021, 2871). In this sense, the ATP represents an inversion of Marcus et al.'s DMM, which prioritizes analogical generalization and falls back on rule application only when the former fails. On the other hand, Veríssimo and Clahsen (2014) describe their Default Generalization Learner (DGL; c.f. §2.3.1) as a dual mechanism model even though it explicitly represents all generalization as rule application. They argue that highly context-sensitive rules are effectively equivalent to analogical generalization, so treating context-free default rules in a qualitatively different manner — i.e. prioritizing them — constitutes a dual mechanism approach. In this understanding, the *level of abstraction* creates two separate mechanisms, rather than the actual mechanism (which in this case is always rule application).

While these models differ on the exact mechanics of dual mechanism realization, all of them share the focus on input conditions characteristic of the rule generation perspective on morphological regularity (Table 2.1). Of course, the minority default analysis requires this focus: an apparently unrestricted distribution over linguistic environments is the key argument for analyzing -*s* as a regular German plural suffix.

3.2.2 Gender-conditioned defaults

An alternative linguistic analysis of German plural inflection posits that, in lieu of one elsewhere-conditioned minority default, there are two gender-conditioned majority default classes: -*en*

---

[2] To be precise, the model is trained on roughly 10,000 nouns from the UniMorph corpus, which falls somewhere between the average vocabulary of German third-graders (5,000 nouns) and eighth-graders (20,000 nouns; Segbers and Schroeder, 2017).

for feminine nouns, and *-e* for nonfeminine nouns. Given the obvious statistical associations between grammatical gender and plural class (Figs. 3.1, 3.2), this idea has been proposed by many linguists over the years. For instance, Bloomfield suggests that ”German plural nouns are derived from singulars by the addition of bound forms which differ according to the gender of the underlying singular: masculine nouns add [-e], with certain vowel-changes [...] feminine nouns add [-en]” (1933, 211). Later research has built on this idea (e.g. August, 1979; Wegener, 1994; Wiese, 1996; Indefrey, 1999), with the most formal theoretical articulation of gender-conditioned default plural classes developed by Bittner (1994, 1999).

Gender conditioning somewhat complicates the concept of “default,” as a gender-restricted input condition is not the Elsewhere Condition, strictly speaking. On the other hand, grammatical gender is a universal property of German nouns; since every noun has a grammatical gender, two rules restricted only by this feature can in principle cover all nouns. This is true especially if nonfeminine nouns are grouped together in the negatively defined context [-feminine], as we see in rule 5 of Table 3.4. In fact, Yang's Tolerance Principle account paradoxically illustrates both the gender-conditioned default and minority default analyses: “the default suffixes — *-(e)n* for [+fem] and *-e* for [-fem] — may achieve productivity for significant subsets of the nouns” (2016, 136). It is not clear how any noun would default to *-s* via rule 6 in this analysis (Table 3.4), if the grammatical gender input condition has already assigned *-en* in rule 1 or *-e* in rule 5.

Gender-conditioned defaults address multiple criteria for morphological regularity (c.f. Table 2.1). In the generative tradition, they are typically formulated as *rules* (e.g. in Table 3.4), and thus defined with respect to their input conditions. Both of the proposed default suffixes appear in a wide range of linguistic contexts, as seen in Table 3.3: *-en* appears in 50 unique feature combinations, second only to *-s*, while *-e* has a highly variable distribution over contexts — $H(F, G) = 4.57$, again second only to *-s*. On the other hand, the reason that *-e* and *-en* appear in so many disparate contexts is largely due to their high *type frequency* (Table 3.1), so they are also consistent with an output-focused view of regularity. Overall, the gender-conditioned default analysis integrates information from both the input and output domains, making it most consistent with the perspective of regularity as *predictability*. The high amount of mutual information between gender and plural class (Table 3.2) makes this approach particularly compatible with the Low Conditional Entropy Conjecture (§2.1.2; Ackerman and Malouf, 2013). If gender-conditioned default classes are the predominant mode of plural generalization, conditional entropy and thus I-complexity should decrease as new words enter the German lexicon. I evaluate how grammatical gender affects plural generalization in Chapters 5 and 6.

3.2.3 Schema theory

Outside of rule-based default analyses, the main alternative linguistic account of German plural inflection is Köpcke's application of schema theory (Köpcke, 1988, 1993, 1998), developed within the broader framework of usage-based linguistics (e.g. Bybee, 1995; Goldberg, 2006; Bybee, 2010) and most directly aligned with the *type frequency* criterion of morphological regularity (§2.1.4). In this approach, inflected wordforms are stored in the lexicon along with the relevant grammatical features (e.g. *Katze*-singular, *Hund*-singular, *Kühe*-plural, *Katzen*-plural, etc.). Frequency is a key driver of linguistic processing in schema theory: a lexical pattern with high type frequency — for example, an *-en* suffix on plural words — may be abstracted as a *schema* (e.g. *-en*-plural). The other key driver is reliability, or cue validity: if a lexical pattern occurs more frequently with one grammatical feature (e.g. if, hypothetically, 90% of the words ending in *-en* are plural and the remaining 10% are singular), then that pattern has higher cue validity for that feature.

This incrementally increasing abstraction and use of cue reliability is similar in spirit to the approach of the Minimal Generalization Learner (MGL, §2.3.1; Albright and Hayes, 2003). The key difference, however, is that schemas represent the association between an output inflected form and its corresponding grammatical feature (e.g. *-en*-plural), rather than the mapping from an input lexeme to an output inflected form; in schema theory terminology, they are "product-oriented" representations, compared to "source-oriented" rules.[3] Schema theory relies crucially upon stored exemplars in memory (especially words with high token frequency; Bittner and Köpcke, 2016), but schematic abstraction over word types makes this approach closer to prototype-based than exemplar-based models, suggesting an alignment with artificial neural networks and discriminative learning models (Chandler, 2017, c.f. §2.3.3).

In terms of morphological regularity criteria (Table 2.1), schema theory has the most clear

---

[3] At least, first-order schemas are product-oriented; researchers have also proposed second-order schemas (e.g. Booij, 2016; Köpcke and Wecker, 2017), which represent rule-like input-output mappings between two first-order schemas (e.g. *-e*-singular<>*-en*-plural).

focus on *type frequency* of output inflected forms, but it is also informed by other criteria. For instance, Bybee (1995) emphasizes that generalization is driven predominantly by type frequency, but also concedes that the concatenativity or "openness" of the rare -*s* plural suffix allows it to generalize more than type frequency alone would predict. As discussed in §2.1.4, it is challenging to formally specify the causal role of type frequency, and how it interacts with other factors relevant to morphological generalization. My impression is that the flexible and data-driven nature of schema theory poses similar barriers to generating clear and testable predictions in the case of German plural inflection, although this possibly reflects my limited understanding rather than broader conceptual issues. Computational models may help by clarifying predictions in specific instances. Bybee (1995) notes that artificial neural network models share many key assumptions of the schema analysis, including the key role of type frequency in generalization — although she maintains they are still too "source-oriented," or focused on the input. In Chapter 7, I evaluate neural network predictions along with other models conceivably related to schema theory, including exemplar (§2.3.3) and Bayesian (§2.3.2) models.

3.3 Behavioral evidence

The previous section reviewed three linguistic theories proposed to account for the observed facts of German plural inflection: minority default theory, gender-conditioned default theory, and schema theory. As discussed in §2.2, some — but not all — theoretical linguistic accounts make behavioral predictions which can be evaluated with psycholinguistic studies. Of the three theories considered here, two have been proposed to account for plural generalization behavior by adult German speakers in wug test experiments. Köpcke (1988) uses schema theory (§3.2.3) to analyze data from a production experiment, and Marcus et al. (1995) use minority default theory in conjunction with the Dual Mechanism Model (§3.2.1) to analyze data from a rating experiment. Although the gender-conditioned default theory (§3.2.2) is prevalent in the linguistic literature, researchers have largely not used it to derive behavioral predictions, though this would certainly be possible.

In this section, I first review the behavioral predictions associated with these three linguistic theories using the analytical framework of regularization (§3.3.1). I then review the behavioral findings of Köpcke (1988) and Marcus et al. (1995) (§3.3.2), with in-depth discussion of the latter as it is the experimental basis for further work in this dissertation. Finally, I provide an overview of behavioral findings from other studies focused on German plural generalization (§3.3.3).

3.3.1 Four ways to generalize

As reviewed in §2.2.2, the regularization literature offers a framework to quantify how lexical patterns influence speaker behavior. Taking the German noun lexicon as a reference distribution, I use this framework to operationalize *four possible outcomes* for how German speakers might generalize plural inflection, based on the three linguistic theories discussed above.

Minority regularization

According to the *minority default* hypothesis, when German speakers encounter novel forms which are dissimilar to the existing lexicon, they should preferentially generalize the plural class -*s* to a greater extent than predicted by its lexical frequency (Table 3.1). In other words, they are expected to overproduce, i.e. regularize, using a minority variant. This behavior is unexpected, but attested nonetheless, in artificial language learning contexts; however, in the case of German plural inflection, overproduction of -*s* can be interpreted as reflecting speakers' knowledge of lexical structure, i.e. the underlying minority default.

Conditional regularization

The *gender-conditioned default* hypothesis predicts that German speakers should rely on grammatical gender when generalizing plural classes, based on the strong statistical association between these two lexical patterns. The *default* part in particular suggests that gender conditioning should be even stronger in the case of atypical novel words, as these are more likely to be assigned the default class. Increasing the statistical interdependence (i.e. mutual information; Table 3.2) between grammatical gender and plural class would decrease I-complexity, consistent with the Low Conditional Entropy Conjecture (Ackerman and Malouf, 2013). This hypothesis could also be realized asymmetrically for different gender categories, as there is stronger statistical evidence for one conditional pattern (i.e. feminine gender predicting -*en*), reflected by a wider consensus for this default relation in the linguistic literature (e.g. Marcus et al., 1995; Wiese, 1996; Yang, 2016; Bittner and Köpcke, 2016). Note, however, that in artificial language learning studies, conditional regularization is typically observed over multiple population-level iterations rather than in spontaneous individual behavior (e.g. Smith and Wonnacott, 2010). Accordingly, the gender-conditioned default hypothesis of German plural inflection has largely developed at the level of linguistic analysis, and has not often informed psycholinguistic behavioral predictions.

General probability-matching

Instead of regularizing, German speakers could also *probability-match* to the overall plural class frequencies observed in the lexicon (Table 3.1). This behavior would align with the predictions of *schema theory*, which posits that the type frequency of a lexical pattern is the strongest factor influencing how it is generalized (Bybee, 1995). It would also align with many artificial language learning studies which find that adult speakers probability-match (Newport, 2016), and some natural language studies that report frequency-matching in morphological generalization (e.g. Hayes et al., 2009).

Conditional probability-matching

The final behavioral possibility is the most open-ended, and compatible to some extent with each considered theory: German speakers may generalize plural inflection in such a way as to *conditionally probability-match* particular lexical distributions. As discussed in §2.2.2, while the artificial language learning literature finds little evidence for spontaneous conditional *regularization*, adult speakers are capable of learning and reproducing conditional *probabilities* (e.g. Hudson Kam, 2015). Artificial language studies can isolate and manipulate conditioning factors in a controlled manner; a natural language lexicon, however, contains many potential conditioning factors, resulting in countless possible distributions to match. So which distributions are relevant to predicting speaker behavior?

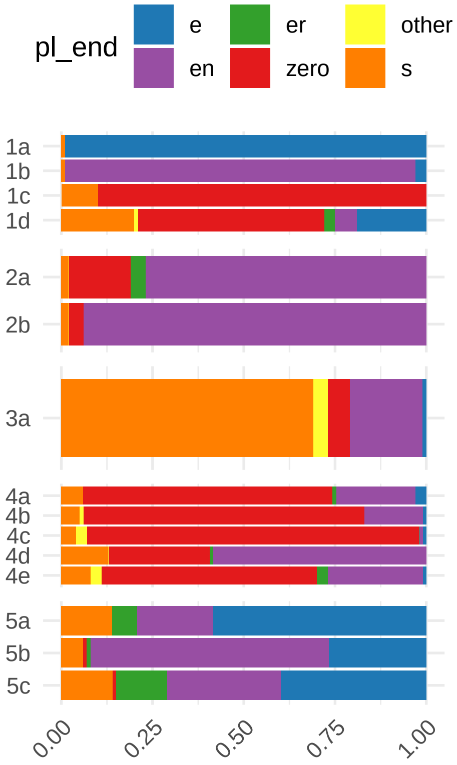
Linguistic theory seeks to characterize the structural properties which organize the lexicon; if we expect lexical structure to influence speaker generalization, then the linguistic accounts of German plural inflection reviewed above could point us to the right conditional factors. In principle, schema theory is the most compatible with any form of probability-matching behavior, but this does not helpfully constrain the hypothesis space. Absent a fully specified model of German plural inflection, schema theory is equally compatible with overall type frequency matching, and, say, gender-conditioned type frequency matching — two strategies which can result in very different patterns of morphological generalization behavior. Gender-conditioned default theory identifies grammatical gender as the key factor structuring plural class organization in the German lexicon, so this account is also in principle compatible with gender-conditioned type frequency matching behavior. Marcus et al.'s dual mechanism (DMM) minority default analysis effectively posits a nested system of probability-matching conditioned on lexical similarity: novel words above a certain threshold of similarity to the existing lexicon undergo analogical generalization, i.e. probability-matching conditioned on lexical similarity, while words below that similarity threshold are instead assigned the minority *-s* class by default. Here again, it is unclear whether grammatical gender is included or excluded when computing lexical similarity.

Summary

Each theory discussed here is compatible with a range of plural generalization behaviors, all of which include some form of conditional probability-matching. The minority default DMM account predicts probability-matching conditional on lexical similarity for words above an unspecified similarity threshold, and minority variant regularization below that threshold. Gender-conditioned default theory predicts conditioning on grammatical gender, which can take the form of either conditional probability-matching or conditional regularization. Schema theory predicts probability-matching either to the overall type frequencies of plural classes or their frequencies conditioned on another factor, such as lexical similarity or grammatical gender. Each linguistic theory of German plural generalization is associated with a wide range of possible behaviors, which complicates evaluation. Computational implementations can be particularly helpful in generating more concrete predictions. We will first review the findings of previous behavioral studies (§3.3.2, §3.3.3), then consider how this phenomenon has been computationally modeled (§3.4).

3.3.2 Köpcke (1988) and Marcus et al. (1995)

In this section, I review behavioral evidence from the wug tests conducted by Köpcke (1988) and Marcus et al. (1995). Köpcke found that speakers generalized near-categorically in response to certain novel noun stimuli, but for other stimuli, they showed gradient patterns consistent with schema theory. Marcus et al.'s experiment used stimuli from the latter category to assess the Dual Mechanism Model (DMM); they report behavioral patterns consistent with the minority default analysis.



Subc.	Gender	Suffix	$H(C)$	% Maj
1a	M	-ling	0.08	99
1b	F	-ung/-schaft	0.27	96
1c	N	-chen	0.47	90
1d	N	-lein	1.88	51
2a	M/N	-e	1.02	77
2b	F	-e	0.38	94
3a	M/N/F	-a/o/u/i	1.33	69
4a	M/N	-el	1.31	68
4b	M/N	-er	1.06	77
4c	M	-en	0.59	91
4d	F	-el	1.41	58
4e	F	-er	1.62	59
5a	M	-	1.59	58
5b	F	-	1.28	65
5c	N	-	1.91	40

(a) Plural productions in each subcategory.

(b) Gender, suffix, entropy over plural class productions $H(C)$, and percent productions of majority class '% Maj' for each subcategory.

Table 3.5: Plural production data, Köpcke (1988, Table 3). Entropy and percent majority calculations conducted by the author.

Köpcke (1988)

Köpcke (1988) conducted an oral wug test in which 40 German students were asked to produce the plural inflected form of 50 novel nouns. The novel noun stimuli fell into five broad categories: 1) ending in a derivational suffix (i.e. *-ling, -schaft, -chen*); 2) ending in a schwa (i.e. *-e*, c.f. Yang's rule 3, Table 3.4); 3) ending in a full vowel (i.e. *-a/o/u/i*); 4) ending in a pseudosuffix (i.e. *-el, -er, -en*, c.f. Yang's rule 2, Table 3.4); and 5) monosyllabic. The data is visualized in Table 3.5, along with entropy measures for the distribution of plural class productions within each subcategory. For the first category with derivational suffixes, speakers typically show near-categorical plural generalization, i.e. 90% or higher convergence on one plural class; for the last category, they show much more variability; and intermediate variability for the categories in between. Köpcke uses schema theory to analyze several of the observed patterns, and concludes that, aside from the input-determined first category, the other four categories show evidence of speakers matching to output plural schemata. Köpcke further considers which suffixes are *over–* or *undergeneralized* by participants (1988, Table 7), relative to 1) their observed frequency in the relevant lexical environments, and 2) the predictions of a generative rule-based plural analysis (Mugdan, 1977). He finds that participants somewhat overgeneralize *-en* (e.g. Category 3), and strongly overgeneralize *-s* (e.g. Categories 1 and 4).

The monosyllabic stimuli in Category 5 are of particular interest for two reasons. Firstly, they received the smallest proportion of zero plural class generalizations from study participants. Köpcke interprets this result to mean that these stimuli have cues associated with singular rather than plural schemata; he identifies the "ideal singular" noun form as one that is a) nonfeminine (masculine or neuter), b) monosyllabic, and c) ending in a consonant stop. Secondly, they received the most consistently variable plural class assignments from participants across all grammatical gender categories, as shown by the entropy and majority class percentage values in Table 3.5. Existing monosyllabic words in the German lexicon show a similar range of observed plural classes. The Duden, the authoritative reference on German grammar, discusses these as *Kernwörter* or "seed words":

> Für Substantive ohne charakterischen Wortausgang (sogenannte Kernwörter) können demgegenüber nur Tendenzen angegeben werden, die es lediglich gestatten, die Pluralendung eines Kernwortes mit hoher Wahrscheinlichkeit vorauszusagen. Letzte Sicherheit ist aber nicht gegeben, sodass in diesem Bereich (ca. 2000 meistens einsilbige Kernwörter) auf die Pluralangaben des Wörterbuchs nicht verzichtet werden kann. (Gelhaus, 1998, 229)
>
> *For nouns without a characteristic word ending (so-called seed words), only tendencies can be identified, which merely permit the prediction of a seed word's plural class with high probability. One cannot be sure, however, so in this area (about 2,000 mostly monosyllabic seed words) a dictionary is needed to identify*

Rhyme	Score	Non-Rhyme	Score
Bral	-7.24	Bnaupf	-12.95
Kach	-6.27	Bneik	-11.97
Klot	-5.68	Bnöhk	-15.08
Mur	-6.13	Fnahf	-13.93
Nuhl	-8.62	Fneik	-11.49
Pind	-7.34	Fnöhk	-15.16
Pisch	-6.59	Pläk	-12.09
Pund	-7.48	Plaupf	-10.07
Raun	-4.70	Pleik	-9.06
Spand	-7.51	Pnähf	-14.55
Vag	-5.43	Pröng	-13.71
Spert	-5.22	Snauk	-11.19
Mean	-6.52	Mean	-12.61

Table 3.6: Rhyme and Non-Rhyme stimuli used in experiments, originally developed by Marcus et al. (1995). Scores report negative log likelihoods assigned to the stimuli by a character n-gram model (Heafield, 2011) trained on the CELEX noun lexicon.

the plural form.[4]

Marcus et al. (1995)

The behavioral experiment conducted by Marcus et al. (1995) builds upon two of Köpcke's key findings: the overgeneralization of *-s*, and the variability of plural class assignments for monosyllabic nouns. The goal of their experiment was to evaluate the minority default analysis of the *-s* plural class. Based on this analysis, they hypothesized that speakers would preferentially generalize *-s* — that is, *overgeneralize* *-s* relative to its expected lexical frequency (following Köpcke's analysis), or *regularize* in artificial language learning terminology (§2.2.2) — in linguistic contexts corresponding to the Elsewhere Distribution (§2.1.3).

Marcus et al. develop stimuli representing three different elsewhere-distributed contexts. The most critical context is phonological. Out of the 24 novel noun stimuli used in this experiment, 12 are Rhymes, meaning they are phonologically typical and rhyme with existing German nouns — while the other 12 Non-Rhymes contain phonotactically valid but rare character sequences. All stimuli are shown in Table 3.6, along with log likelihood scores computed by a 5-gram character model with backoff (Heafield, 2011) trained on the CELEX noun lexicon. As expected, all of the Rhyme stimuli have higher likelihood under the character language model than all of the Non-Rhyme stimuli. Note as well that all of the stimuli are monosyllabic,

---

[4] This and subsequent translations are by the dissertation author, unless otherwise indicated.

Condition	Examples (English translation)
Root	I took a green X for my cold. But the white $X.pl$ are often cheaper and work better.
Borrowing	The French X looks best in black. But $X.pl$ actually look good in every color.
Name	My friend Hans X and his wife Helga X are a bit strange. The $X.pl$ always try to put on their shoes before their socks.

Table 3.7: Marcus et al. (1995) study design.

and therefore drawn from the part of the German noun lexicon with the highest variability in plural class assignment, as discussed above. Following the logic of the Dual Mechanism Model (DMM, §3.2.1), Marcus et al. reason that Rhymes would be sufficiently similar to existing German nouns to undergo analogical generalization, resulting in the assignment of a non--*s* plural class; the dissimilarity of Non-Rhymes, however, would prevent analogical generalization to some extent, leading speakers to fall back on the minority default -*s*. Therefore, despite the high variability in plural class generalization observed for nouns of this type, Marcus et al. predict that speakers will generalize -*s* significantly more to Non-Rhymes than to Rhymes.

The other two elsewhere contexts are semantic, and conveyed by the content of the sentence surrounding the novel word. Marcus et al. present the 24 novel nouns (Table 3.6) in three different sentence conditions: as standard German root words, as loanwords borrowed from another language, and as proper names. Translated examples of each condition are presented in Table 3.7. They consider both the loanword and proper name contexts as elsewhere conditions, so predict greater -*s* generalization in both conditions. Proper names obligatorily take -*s* in German (except when phonologically blocked, such as names that already end in -*s*), and borrowings often take -*s*, so -*s* generalization in both contexts is expected based on the existing German lexicon. This is not the case for phonological typicality, so the Non-Rhymes constitute the key test for the minority default hypothesis.

48 adult German speakers were presented the 24 words as a paper-and-pencil rating task. Each word was shown first in its singular form in the context of a sentence, and then speakers were asked to rate each of its eight possible plural forms if umlaut was available (i.e. if the noun had a back vowel), or five forms if umlaut was unavailable; Table 3.8 presents an example rating task. Each word appeared an equal number of times with masculine and feminine gender, and in the three sentence contexts, counterbalanced across participants.

Die Französische KACH sieht schwarz am besten aus.
Aber eigentlich sehen KACH in jeder Farbe gut aus.
Aber eigentlich sehen KÄCH in jeder Farbe gut aus.
Aber eigentlich sehen KACHE in jeder Farbe gut aus.
Aber eigentlich sehen KÄCHE in jeder Farbe gut aus.
Aber eigentlich sehen KACHER in jeder Farbe gut aus.
Aber eigentlich sehen KÄCHER in jeder Farbe gut aus.
Aber eigentlich sehen KACHEN in jeder Farbe gut aus.
Aber eigentlich sehen KACHS in jeder Farbe gut aus.

Table 3.8: Marcus et al. (1995) example rating task for *Kach* in the Borrowing condition.

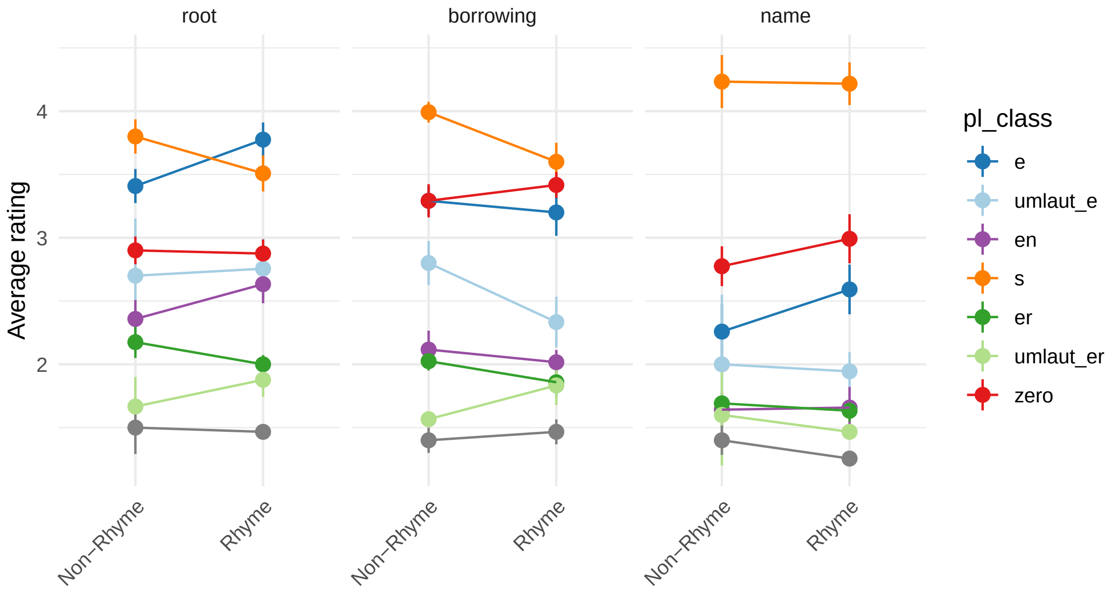


Figure 3.3: Average item-level ratings reported by Marcus et al. (1995).

Marcus et al. (1995) report findings consistent with their minority default hypothesis: participants rated *-s* plural forms higher for Non-Rhyme words than for Rhymes in both the Root and Borrowing conditions. In the Root condition, *-s* was the top-rated plural form for 2 out of 12 Rhyme words, and 7 out of 12 Non-Rhyme words. As expected, participants assigned the highest rating to *-s* forms of all nouns in the Name condition, as *-s* is the obligatory plural class form for German proper names. No significant effect or interaction with gender was reported. Marcus et al. (1995) report averaged ratings for each item and plural variant, which are visualized in Figure 3.3. Clahsen summarizes the results: "We found that the *-s* plural is rated significantly better for nonrhymes than for rhymes, whereas all other plurals produced the reverse pattern. [...] This shows that adults do indeed generalize *-s* plurals to nonrhyming real words and to foreign words in German" (1999*a*, 1049).

Zaretsky and Lange (2016) used the same novel noun stimuli (Table 3.6) in a large-scale replication study with 585 participants and a modified task — written production of the plural noun form given its singular form with a indefinite article[5] (e.g. "ein Kach"), rather than Marcus et al.'s rating task with differing sentence contexts. While they found the same asymmetry in *-s* production favoring Non-Rhymes over Rhymes, two other findings cast doubt on the minority default account. First, the plural class *-en* showed a similar pattern of favoring Non-Rhymes over Rhymes. Second and most importantly, while Marcus et al. found that *-s* was the preferred class for Non-Rhymes in all three different rating categories (c.f. Figure 3.3), Zaretsky and Lange found that participants produced *-e* and *-en* plural classes much more frequently than *-s* for both Non-Rhymes and Rhymes. They conclude that *-s* has no special status: *-e*, *-en*, and *-s* are all productive in modern German, and the apparent *-s* preference found by Marcus et al. is an artifact reflecting the rating rather than production task. However, in a separate large scale (219 participants) written production study using different novel stimuli, Molloy (2018) finds a strong preference for *-s* generalization and concludes in support of the minority default hypothesis.

---

[5] While the German definite articles mark three distinct grammatical genders (*der* masculine, *die* feminine, *das* neuter), the indefinite article distinguishes only nonfeminine *ein* from feminine *eine*.

Summary

In summary, the studies reviewed here present a mixed picture of *-s* generalization. Köpcke (1988) finds that German speakers overgeneralize *-s* relative to its expected frequency in particular lexical environments, and Marcus et al. (1995) find that speakers consistently rate the *-s* plural forms of phonologically atypical nouns (Non-Rhymes) highly relative to phonologically typical nouns (Rhymes). Zaretsky and Lange (2016) find that speakers also produce *-s* plural forms more often for Non-Rhymes relative to Rhymes, but they produce the frequent plural classes *-e* and *-en* more often than *-s* in all categories; however, the findings of Molloy (2018) appear to contradict this. None of the reviewed studies provide data at the level of individual speakers, which would be necessary to categorize the different behavioral possibilities outlined in §3.3.1. I conduct independent replication studies using Marcus et al.'s stimuli in Chapters 4, 5, and 6. Chapter 4 directly investigates the question of *-s*-generalization, and Chapter 6 analyzes the behavior of individual speakers.

3.3.3 Other behavioral studies

In addition to the studies reviewed above focusing on *-s*, several other behavioral experiments on plural generalization have been conducted with adult German speakers. These studies generally have also not released or described their data in sufficient granularity to support analysis at the level of individual speakers (§3.3.1), so we will review them only briefly here.

Several researchers have found evidence for systematic behavior in certain lexical contexts, consistent with the results reported by Köpcke (1988). Gawlitzek-Maiwald (1994) reports similarly systematic generalization of the plural classes associated with derivational suffixes. Independently of elsewhere contexts, *-s* is consistently preferred for nouns ending in a full vowel (Mugdan, 1977; Gawlitzek-Maiwald, 1994; Molloy, 2018). German speakers also reliably generalize *-(e)n* to nouns ending in schwa (Mugdan, 1977; Spreng, 2004, c.f. Molloy, 2018).

Outside of these systematic tendencies, however, most participant responses evince a high degree of variability in plural form generalization. This is apparent on close inspection of the data from these studies, although the variability is often obscured by researchers' focus on aggregate statistics and predictable patterns — reporting practices that make it difficult to identify and measure probability-matching behavior (c.f. Pierrehumbert, 2022, quoted in

§2.2.1). A notable exception to this trend is found in the work of Mugdan (1977), who developed an extensive generative rule-based analysis of German plural inflection, and then concluded that his model was effectively disproved (for psycholinguistic purposes at least) by the variable behavior of participants in his wug test study:

> Die Versuchsauswertung läßt nicht den Schluß zu, das das Verhalten der getesteten Personen bei der Bildung des Plurals unbekannter Substantive in einfacher Weise durch Regeln beschreibbar ist. (Darauf deuten die unterschiedlichen Reaktionon auf ähnliche Testwörter.) Einige Vpn suchten nach analogen Beispielen, andere gestanden, bei den Antworten wild geraten zu haben. [...] Dabei kam es auch zu Bildungen, die es nach den Regeln des Deutschen überhaupt nicht gibt. (Mugdan, 1977, 172)
>
> *The experimental results do not support the conclusion that participants' plural inflection behavior is easily described by rules. (This is evident in the differing responses to similar test items.) Some participants looked for analogical examples, while others appeared to make wild guesses. [...] This even led to plural forms which don't exist under the rules of German.*

Chapters 6 and 7 quantify the variability observed in behavioral experiments, using the regularization framework developed in the artificial language learning literature (§2.2.2, §3.3.1).

Despite its long pedigree in theoretical linguistic accounts, grammatical gender has received relatively little attention in behavioral wug tests, and existing reports remain inconclusive. Some researchers comment upon observed effects of gender, but do not quantify the effects with statistical analysis (Köpcke, 1988; Gawlitzek-Maiwald, 1994; Spreng, 2004; Molloy, 2018). Marcus et al. (1995) report no significant effect or interaction of grammatical gender, but Zaretsky and Lange (2016) find significant and reliable effects of gender on the same stimuli. Finally, Mugdan remarks upon the bewildering *absence* of grammatical gender effects in his study, which was apparently even brought to his attention by participants:

> Es scheint, daß es vor allem der Wortklang war, der die Vpn bei der Auswahl eines Pluralallomorphs leitete, während insbesondere das Genus offenbar weitgehend unbeachtet blieb. (Das wurde auch von manchen Vpn nach Abschluß des Tests erwähnt.) (1977, 172)

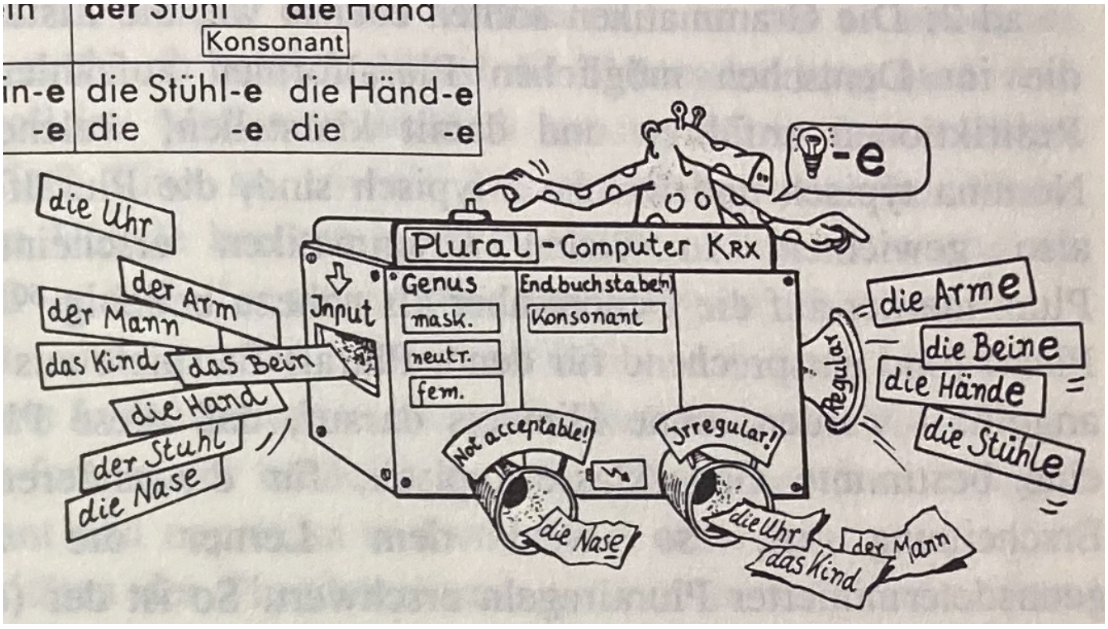


Figure 3.4: The Plural-Computer, which assigns German nouns a regular plural form based on their gender (Genus) and final character/s (Endbuchstabe/n) (Mugdan, 1977, 53).

> *It seems that the participants' selection of plural allomorphs was guided above all by the sound of the word, while gender remained largely ignored. (Several participants even mentioned this at the end of the test.)*

Chapters 5 and 6 investigate the behavioral effects of grammatical gender.

3.4 Computational evidence

The previous sections have reviewed three linguistic theories of German plural inflection — minority default (§3.2.1), gender-conditioned defaults (§3.2.2), and schema theory (§3.2.3) — which are respectively associated with the rule generation, predictability, and type frequency views of morphological regularity (§2.1.1). As discussed in §2.3, various computational implementations have been used to model morphological generalization, reflecting these different perspectives. In this section, I review computational approaches to modeling German plural inflection, using the task as framed by Rumelhart and McClelland (1986): to learn a mapping from singular to plural inflected forms given the German noun lexicon (c.f. Figure 3.4), and use this mapping to predict speaker generalization behavior.

In contrast to the extensive theoretical literature in this area, computational modeling research on German plural inflection has proven roughly as inconclusive as the behavioral research reviewed above. The modeling debate has historically been dominated by the influential Dual Mechanism Model (DMM; Marcus et al., 1995; Clahsen, 1999*b*) and the associated challenge of modeling a minority default (§3.2.1), with the result that modeling research in this area has often been framed as 'single- vs. dual-route' instead of broadly focused on morphological regularity criteria and model capacities as described in §2.3. Moreover, a lack of fine-grained behavioral data on German plural generalization (c.f. §3.3) has led many modeling papers to evaluate generalization via a held-out test set of plural forms from the existing lexicon, rather than direct comparison to novel inflected forms produced by speakers. This gap is especially apparent for models which treat inflection as a classification rather than sequence generation task, which is true of nearly all models reviewed here. Pinker and Ullman characterize analogical classifiers as dependent on rules for sequence generation, and hence inherently dual mechanism:

> To convert the choice into an actual form, some other mechanism would have to copy the stem and apply the pattern corresponding to the selected unit. Such a mechanism is simply a rule. (2002, 458)

Regardless of the rule debate, classification models are clearly not capable of treating inflection as an open-ended sequence generation task, which necessarily limits their capacity to represent speaker behavior; no classifier could generate the "plural forms which don't exist under the rules of German" (1977, 172) which Mugdan observed from his participants.

Despite these issues, computational modeling studies have provided key insights into the nature and complexity of German plural inflection. I structure this review of computational models in relation to the three theories considered above, and the morphological phenomena they seek to account for: minority default generalization of *-s* (§3.2.1), gender-conditioned generalization of the majority plural classes *-en* and *-e* (§3.2.2), and the integrated influences of type frequency and lexical cues in plural generalization broadly (§3.2.3). Some models considered here represent the relevant phenomena by design, while others use general learning principles to arrive at the target generalization behavior.

Minority default learners

A core argument of the DMM is that dual-route models are necessary to account for the observed generalization behavior of the -*s* plural class in German (Marcus et al., 1995). This claim has been tested by Nakisa and Hahn (1996); Hahn and Nakisa (2000), who compare DMM implementations with single-route models using analogical generalization. Nakisa and Hahn (1996) train an exemplar model (the GCM; Nosofsky, 1988) and a three-layer feed-forward neural network classifier on the German noun lexicon, using phonological representation only (i.e. no grammatical gender). They additionally implement dual-route versions of both models by excluding nouns which take the -*s* plural from training, and instead assigning -*s* by default to any inputs below a learned similarity threshold. Nakisa and Hahn (1996) evaluate the models on generalization to a held-out test set, and find that the dual route models never outperform their single-route counterparts, with the neural network model achieving the highest overall generalization accuracy. Hahn and Nakisa (2000) build on these results with comparison to speaker production data on the stimuli developed by Marcus et al. (1995). The behavioral data is not analyzed in detail, but they report that the predictions of the exemplar-based GCM show a closer overall fit to speaker productions than the dual-route version of the GCM. Overall, these modeling results indicate no clear advantage for DMMs in German plural generalization.

Subsequent work has shown that the minority default -*s* can be learned by very different single-route model architectures. Feldman (2005) trains ten three-layer neural classifiers on a dataset of Austrian German child-directed speech, and tests their default activation on input vectors with zero-valued features. She finds evidence for default activation of -*en* and -*s* when no grammatical gender information was provided, and for -*en* as the default for feminine nouns; however, she also finds unexpected evidence for default -*s* generalization for masculine nouns, which raises the possibility of divergent results based on a non-standard variety of German. Belth et al. (2021) train the symbolic ATP (§2.3.1) on samples of standard German child-directed speech ranging from 100 to 400 nouns, and report that the ATP learns a minority default -*s* rule in roughly 20% of the training simulations. These findings from disparate model classes show that the dedicated hybrid architecture of DMMs is not necessary to learn minority default *-s* generalization; however, they do not conclusively establish whether doing so supports speaker-like behavior (c.f. Chapter 4).

Gender-conditioned default learners

Gender-conditioned plural generalization is less challenging for statistical learners compared to minority default generalization, in light of the robust statistical association between grammatical gender and plural class in the German noun lexicon. Symbolic statistical models can readily learn gender-conditioned defaults, as shown by the example of Yang's Tolerance Principle analysis (§3.2.2, Table 3.4). As discussed in §2.3.3, neural network models are generally capable of learning complex conditional relations, and Goebel and Indefrey (2000) find that this capacity holds for the gender-plural association. They train five recurrent neural network models with short-term memory on a frequency-weighted sample of nouns produced by six-year-old German children, and test their models on the stimuli developed by Marcus et al. (1995). Their most robust finding is gender-conditioned application of the *-e* and *-en* plural classes to both Rhymes and Non-Rhymes, which they interpret as morphologically regular generalization (see also Indefrey, 1999):

> We must, therefore, conclude that the German -e and -n plural are not irregular. The German plural system consists of two frequent regular allomorphs (-e and -n), two infrequent irregular morphs and umlaut) and one infrequent regular allomorph (-s). All regular allomorphs phonologically are unrestricted and can serve as defaults. (Goebel and Indefrey, 2000, 193)

While Goebel and Indefrey's neural model *learned* to condition on grammatical gender, Williams et al. (2020) explicitly condition on gender in a recent analysis of German noun inflection. They use recurrent neural networks to estimate the mutual information between the inflection class (including plural number and also case marking) and the phonological form and semantic meaning of German nouns. Due to the high mutual information between gender and inflection class across the lexicon, the authors condition all other estimated values on grammatical gender. They find that form and meaning contribute independently to the predictability of inflection classes. Their computational implementation reflects theoretical linguistic analyses which posit grammatical gender as the primary factor organizing the German noun lexicon (§3.2.2). More recent work by Dankers et al. (2021) establishes that neural models continue to learn gender-based plural generalization. I consider the influence of grammatical gender on predictions from neural models in Chapters 4 and 5, and from a broad spectrum of computational models in Chapter 7.

Associative and predictive learners

While the preceding sections focus on computational modeling of behaviors targeting specific plural inflection classes — namely, minority *-s* generalization, and gender-conditioned generalization of *-e* and *-en* — several researchers have explored applications of analogical modeling (i.e. exemplar-based and neural network models; §2.3.3) to German plural generalization more broadly. As discussed above, many of these approaches rely on held-out test sets for evaluation. Wulf (2002) applies Analogical Modeling (Skousen, 1989) to a subset of the CELEX German noun lexicon (700 nouns) and achieves 72% test set accuracy. Daelemans (2002) trains on a larger subset of the lexicon (20,000 nouns) and finds 86.6% test accuracy from AM, but this is bested by a Memory-Based Learner (i.e. TiMBL, closer to a nearest-neighbor model) which achieves 89.7% accuracy. As discussed above, Hahn and Nakisa (2000) find that a three-layer neural network model outperforms another exemplar model (the GCM), achieving 82.7% accuracy when trained on 7,000 nouns. These analyses indicate that exemplar-based and neural models can achieve some degree of successful generalization within the existing German noun lexicon; however, they do not directly relate to schema theory accounts of speaker generalization behavior, such as Köpcke's (1988) proposed set of cues distinguishing singular and plural schemas. The reader is referred to Heitmeier et al. (2021) for a comprehensive and up-to-date discussion of semantic and phonological cues in discriminative modeling of German plural inflection, including evaluation on behavioral data (McCurdy, Goldwater and Lopez, 2020, c.f. Ch. 4). Chapter 7 evaluates neural and exemplar models, inter alia, in terms of their fit to speaker behavior.

3.5 Conclusion: Research Questions

This chapter has reviewed how these perspectives have been applied to the specific domain of German plural inflection. We have considered three linguistic theories of German plural inflection — minority default (§3.2.1), gender-conditioned default (§3.2.2), and schema theory (§3.2.3) — and their respective relation to the rule generation, predictability, and type frequency perspectives of morphological regularity. We have also considered the behavior predicted by each of these accounts, and the evidence from existing behavioral studies of German plural generalization (§3.3). Finally, we have reviewed computational modeling studies of German plural inflection, and considered how they align with, or diverge from, theoretical and behavioral findings (§3.4).

The literature reviewed here has informed the two key research questions of this dissertation: how do German speakers generalize plural inflection, and can computational models learn to generalize similarly? Subsequent chapters address these questions in light of the theories and issues raised above. Chapter 4 reports a behavioral experiment to test the minority default hypothesis proposed by Marcus et al. (1995), and uses the resulting data to evaluate neural and symbolic models of morphological generalization. Chapter 5 investigates the role of grammatical gender in two behavioral experiments, and Chapter 6 builds on this work with an additional experiment and comparative analysis to lexical distributions. Chapter 7 gathers all the data from the four behavioral experiments reported in previous chapters, and uses it to evaluate a range of computational models — neural, symbolic, exemplar, and Bayesian — as well as baselines reflecting a range of lexical distributions.

Chapter 4

Speakers default to variation

4.1 Introduction

This chapter presents the first of four behavioral experiments in this dissertation, and several related computational modeling experiments. McCurdy, Goldwater and Lopez (2020, §4.2) report behavioral Study 1 and a modeling experiment with the neural Encoder-Decoder (ED; Kann and Schütze, 2016) proposed as a cognitive model of morphological generalization by Kirov and Cotterell (2018). §4.3 presents additional modeling experiments and analysis which ultimately contradict the conclusions drawn in §4.2. I include here the contents of McCurdy, Goldwater and Lopez (2020), lightly edited, because the publication has already been cited and built upon by other researchers (e.g. Beser, 2021; Belth et al., 2021; Dankers et al., 2021; Heitmeier et al., 2021). In lieu of presenting an entirely revised analysis, I clarify the updated results and discuss their implications at some length (§4.4).

The goal of this chapter is to test two influential claims about morphological generalization advanced by Marcus et al. (1995). The first claim is behavioral. In a rating-based wug test (§3.3.2), Marcus et al. find that German speakers prefer *-s* plural forms for phonologically atypical Non-Rhyme nouns (e.g. *Bneik*) compared to nouns which sound like existing German words (e.g. *Bral*). They interpret this result in support of the **rule generation** criterion for morphological regularity (§2.1.1): because the rare suffix *-s* has an unrestricted *elsewhere* distribution over input features (§2.1.3), it is a regular inflection class, and therefore preferentially generalized to atypical inputs (§3.2.1). In Study 1 (§4.2), compared to the original Marcus et al. results, I find much lower use of *-s* in both production and rating tasks, but reproduce speakers' statistically significant Non-Rhyme *-s* preference. Subsequently (§4.3.2), however, my analysis leads me to interpret this effect as a positively-conditioned response to confounded Non-Rhyme stimuli ending in *-k*, rather than a negatively-conditioned response to default circumstances, i.e. phonological atypicality. I conclude that the behavioral evidence does not support a rule generation analysis of *-s*.

The second claim is computational. Marcus et al. argue that speakers' Non-Rhyme *-s* preference is readily modeled as a symbolic default rule (§2.1.3), and therefore learnable for statistical models with a symbolic component, either rule-based learners (§2.3.1) or hybrid Dual Mechanism Models (DMMs; §3.2.1, §3.4). Statistical learners *without* rules, however — such as neural network models — lack an explicit representational mechanism for a negatively conditioned input criterion (i.e. default rule), relying instead upon the **type frequency** criterion for regularity (§2.1.1). Given an atypical input, these models should preferentially generalize frequent classes such as *-e* or *-en* over the infrequent *-s*. Though other researchers have demonstrated that single-mechanism models, including neural models, can learn to treat *-s* as a default (§3.4), none have directly compared model and speaker productions.

I evaluate the computational claim by training multiple instances of a neural Encoder-Decoder model (ED; Kann and Schütze, 2016) and a rule-based symbolic model (ATP; Belth et al., 2021) on the same German lexical resource (UniMorph; Sylak-Glassman et al., 2015). As reported by McCurdy, Goldwater and Lopez (2020), I find that the ED model fails to reproduce the Non-Rhyme *-s* preference when trained to predict plural forms using the singular noun's orthographic wordform *and* grammatical gender (e.g. *das Bral*). In subsequent analysis (§4.3.1), however, I find that both the ED and ATP reproduce the Non-Rhyme *-s* preference when trained on wordform alone, without grammatical gender (e.g. *Bral*). Moreover, both models generalize *-s* more often to Non-Rhymes not due to learning a default rule, but due to a stimulus confound: half of the Non-Rhyme stimuli end in *-k*, which has a higher rate of lexical co-occurrence with the *-s* plural compared to other word-final segments. This finding leads me to reanalyze the speaker data and attribute the behavioral effect to the same confound.

Finally, though both computational models capture the Non-Rhyme *-s* preference, I find that they differ from speaker behavior in another critical respect. Speakers consistently generalize a variable *distribution* of plural classes to each individual stimulus item — a distribution which, incidentally, looks quite close to those classes' lexical type frequencies — but both the neural and symbolic model are *overconfident* in comparison, strongly preferring one plural class for a given item. Marcus et al.'s criticism is correct in a sense: statistical learners over-rely on the typically frequent class with the highest conditional likelihood, while speakers appear to default to variation for both typical and atypical .

4.2 Study 1

The material in this section has been published as McCurdy, Goldwater and Lopez (2020).

4.2.1 Introduction

Morphology has historically been the site of vigorous debate on the capacity of neural models to capture human speaker behavior, and hence ground claims about speaker cognition. In 1986, Rumelhart and McClelland described a neural network model which learned to map English present tense verbs to their past tense forms. Importantly, the network handled both **regular** verbs, whose past tense is formed systematically by adding the suffix /-(e)d/ (e.g. *jumped*), and **irregular** verbs where the present and past tenses bear no systematic relationship (e.g. *ran*). The authors suggested their model provided "an alternative [...] to the implicit knowledge of rules" (1986, 218), a claim which sparked considerable controversy. Pinker and Prince (1988) highlighted many empirical inadequacies of the Rumelhart and McClelland model, and argued that these failures stemmed from "central features of connectionist ideology" and would persist in any neural network model lacking a symbolic processing component.

Recently, however, Kirov and Cotterell (2018, henceforth K&C) revisited the English past tense debate and showed that modern recurrent neural networks with encoder-decoder (ED) architectures overcome many of the empirical limitations of earlier neural models. Their ED model successfully learns to generalize the regular past tense suffix /-(e)d/, achieving near-ceiling accuracy on held-out test data. Moreover, its errors result from overapplication of the regular past tense (e.g. *throw–throwed*)—a type of error observed in human language learners as well—as opposed to the unattested forms produced by Rumelhart and McClelland's model. K&C conclude that modern neural networks can learn human-like behavior for English past tense without recourse to explicit symbolic structure, and invite researchers to move beyond the 'rules' debate, asking instead whether the learner correctly generalizes to a range of novel inputs, and whether its errors (and other behavior) are human-like.

This challenge was first taken up by Corkery et al. (2019), who showed that, on novel English-like words designed to elicit some irregular generalizations from humans, the ED model's predictions do not closely match the human data. While these results suggest possible problems with the ED model, English may not be the best test case to fully understand these, since the sole regular inflectional class is also by far the most frequent. In contrast, many languages have multiple inflectional classes which can act 'regular' under various conditions (Seidenberg and Plaut, 2014; Clahsen, 2016).

In this paper, we examine German number inflection, which has been identified as a crucial test case for connectionist modeling (Köpcke, 1988; Bybee, 1995; Marcus et al., 1995; Clahsen, 1999*b*). The German plural system features eight plural markers (c.f. Table 3.1), none of which hold a numerical majority in type or token frequency. Different linguistic environments favor different plural markers (e.g. Köpcke, 1988; Wiese, 1996; Yang, 2016), and even the famously rare suffix *-s* is nonetheless **productive**, in the sense that speakers readily extend it to new words.[1] In their analysis of the German plural system, Marcus et al. (1995, henceforth M95) argue that neural networks generalize the most frequent patterns to unfamiliar inputs, and thus struggle to represent productive but rare classes such as *-s*. We investigate that claim using the novel German-like nouns M95 developed.

Because the design and results of previous human studies have been somewhat inconsistent, and because we want to compare to fine-grained results from individuals (not just published averages), we first collect a new dataset of plural productions and ratings from German speakers. Our speaker data show high variability: no class holds a majority overall, and two less frequent suffixes show a relative preference for phonologically atypical inputs ("Non-Rhymes"). We then compare our human data with the predictions of the encoder-decoder (ED) model proposed by K&C. While our human data paint a more complex picture of the German plural system than M95 claimed, nevertheless M95's central idea is borne out: when given Non-Rhymes, the ED model prefers the most frequent plural class, but speakers behave differently. This finding reveals that while modern neural models are far more powerful than earlier ones, they still have limitations as models of cognition in contexts like German number inflection,

---

[1] For example, the Institut für Deutsche Sprache (`https://www.owid.de/service/stichwortlisten/neo_neuste`) officially added multiple *-s*-inflecting nouns to the German language in 2019, including *Verhütungsapp*, *Morphsuit* and *Onesie*.

where no class holds a majority. The model may correctly identify the most frequent class, but fails to learn the conditions under which minority classes are productive for speakers.

4.2.2 Study 1a: Speaker plural inflection

To evaluate whether neural models generalize correctly, we need to compare their behavior with that of humans on the same task. Unfortunately, no existing datasets were suitable, so our first study asks how German speakers inflect novel nouns.

Background The task of inflecting novel (**nonce**) words is known as the *wug test* (Berko, 1958), and is the standard method to assess morphological generalization in psycholinguistic research (§2.2.1). In this study, we use the wug test to investigate how German speakers generalize plural classes. See §3.1 for a general overview of the German plural system.

As discussed in §3.1 (e.g. Figure 3.1, Table 3.3), the infrequent class *-s* occurs across a wide and diverse range of linguistic environments. For this reason, M95 argue that *-s* is the **minority default** plural, and can apply to any noun regardless of its form (see discussion in §3.2.1 and §3.3.2). M95 claim that generalizing *-s* should be particularly difficult for connectionist, i.e. neural, models to learn: *-s* cannot be generalized based on its frequency, as it is rare, and it cannot be generalized based on similar inputs, as it applies to heterogeneous, unfamiliar inputs.

To assess whether German speakers treat *-s* as a productive default for novel words, M95 developed a list of 24 monosyllabic nonce nouns for wug testing (Table 3.6). The stimuli represented two phonological classes: 'familiar' or Rhyme words, which rhymed with one or more existing words in German (e.g. *Bral*, rhyming with *Fall*; *Spert*, rhyming with *Wert*), and 'unfamiliar' or Non-Rhyme words (e.g. *Plaupf*, *Fnöhk*), which were constructed using rare but phonotactically valid phone sequences. They hypothesized that Non-Rhymes, as phonologically atypical words, should be more likely to take the *-s* plural. See §3.3.2 for further details of the study design. M95 found that *-s* was the top-rated plural form for 2 out of 12 rhyme words, and 7 out of 12 non-rhyme words; while *-e* was rated highest overall, *-s* was the only marker favored more for non-rhymes. Clahsen (1999*a*) cites this asymmetry as crucial evidence for *-s* as the only default plural form, at least with respect to these stimuli.

These results, however, have been called into question. Zaretsky and Lange (2016, henceforth Z&L) conducted a large-scale follow-up study with 585 participants, using the same nonce words but a different task: instead of rating the plural forms within a sentence context, subjects were presented with the noun in isolation (e.g. *Der Bral*) and asked to produce its plural form.[2] They found a much lower preference for *-s* than expected based on M95's results, and a significant effect for feminine (*die*) versus non-feminine (*der, das*) grammatical gender, where M95 did not report an effect of gender. The authors conclude from their data that *-en*, *-e*, and *-s* are all productive in German, and also speculate that task differences (production versus rating) could account for the discrepancy between the two studies.

Data collection

Motivation Although M95 published average rating data for each word in the appendix to their paper, we felt it necessary to collect our own data. Z&L's findings suggest that the M95 *-s* effect might reflect task artefacts: speaker behavior could differ for production and rating tasks, and with and without sentential context for the nonce words. We seek to evaluate K&C's performance claims for ED models, which were based on speaker production probabilities rather than ratings. To do so, we need speaker data which closely parallels the model task: given a noun in isolation, produce its plural inflected form. We collect production data, and also ratings, to see whether speaker behavior is consistent across tasks.

Another issue raised by Z&L's findings is the role of grammatical gender. Although Z&L reported significant gender effects, M95 did not: their reported rating averages combine all gender presentations (e.g. *Der Bral, Die Bral, Das Bral*). Previous experiments have found neural models of German plurals to be sensitive to grammatical gender (Goebel and Indefrey, 2000); therefore, the stimuli presented to speakers should be consistent with model inputs to enable valid comparison. For simplicity, we opted to select one grammatical gender for presentation: neuter, or *Das*. Based on similar experimentation by **?**, speakers do not have a strong majority class preference for neuter monosyllablic nouns, hence this environment may be the most challenging for a neural model to learn. For this reason, we present all stimuli as neuter to study participants.

Method The current study uses the same Rhyme and Non-Rhyme stimuli from M95's original experiment. We collected both production and rating data on plural inflection for the 24 M95

---

[2] Z&L's data is unfortunately not freely available.

nonce nouns through an online survey with 150 native German-speaking participants. Survey respondents were first prompted to produce a plural-inflected form for each noun (i.e. filling in the blank: "*Das Bral, Die* ______").[3] After producing plural forms for all nouns, they were prompted to rate the acceptability of each potential plural form for each noun on a 1-5 Likert scale, where 5 means most acceptable. For example, a participant would see *Das Bral*, and then give an acceptability rating for each of the following plural forms: *Bral, Bräl, Brale, Bräle, Bralen, Braler, Bräler, Brals*.

Survey procedure We designed an online survey comprising three sections, in order of presentation: 1) an introductory production task with existing German words, 2) a nonce-word production task, and 3) a nonce-word rating task. For the introductory production task, eight existing German nouns were used, one from each of the eight plural classes under consideration. The goal of this section was to familiarize participants with the task of producing the plural, and avoid biasing them toward any particular plural marker by showing all eight options. We also hoped that inflecting nouns in Modern High German would encourage participants to approach the following tasks with the standard variety primed, thus reducing the possible effects of dialectal variation. For the second and third sections, the production and rating tasks, the twenty-four M95 nonce words were presented. All stimuli were presented with neuter grammatical gender in the nominative case. In all tasks, each noun was preceded by the article *Das*, indicating neuter gender and singular number, and each prompt for participant responses was preceded by *Die...*, to indicate plural number. The eight existing nouns in the introductory production task were also neuter gender, so followed this pattern as well.

We recruited 192 participants through the online survey platform Prolific[4], using the site's demographic filters to target native German speakers. Participants were additionally asked about their age and exposure to languages other than German within the survey. Participants were shown the three tasks, introduction, production, and rating, in order, meaning that participants had to produce a plural form for all 24 nonce words before performing the rating task. For the production task, participants saw the noun on its own, preceded by *Das*, e.g. *Das Bral*. Above the response box, the text *Die...* appeared, to indicate that a plural form of the

---

[3] The article *das* indicates singular number, neuter gender; as all nouns were presented in neuter gender (see preceding discussion), all nouns were preceded by *das*. *Die* here indicates plural number, so the following noun will be pluralized.

[4] `http://www.prolific.com`

Plural		Prod %	N	Rating (SE)
/-e/	R	**45.3**	815	3.53 (.021)
	NR	**44.7**	805	3.51 (.024)
/-(e)n/	R	25.0	450	**3.73** (.026)
	NR	34.7	624	**3.84** (.025)
/-er/	R	17.4	314	3.08 (.022)
	NR	6.7	120	3.06 (.024)
/-s/	R	4.2	75	2.39 (.027)
	NR	6.4	116	2.52 (.028)
ø	R	2.7	48	2.24 (.020)
	NR	2.7	48	2.38 (.024)
other	R	5.4	98	
	NR	4.8	87	
overall	R		1800	2.99 (.011)
	NR		1800	3.04 (.012)

Table 4.1: Survey results. Production reported as percentages out of all Rhymes (R) and Non-Rhymes (NR); ratings are averages over a 1 (worst) – 5 (best) scale, with standard errors in parentheses. Highest numbers in each category are bolded.

noun should be typed into the response box below the text. For the rating task, participants were prompted to rate each potential plural on a Likert scale of *Sehr gut* ('very good'; 5) to *Sehr schlecht* ('very bad'; 1). After filtering out, data from 150 participants was available for analysis. The cleaned, anonymized survey data will be published online along with this paper.

Results

Our study results are shown in Table 4.1. The production data collected in our survey appears broadly consistent with the distribution observed by Z&L and Köpcke: *-e* is favored in production, followed by *-en*. The rhyme vs non-rhyme comparison is also consistent with Z&L's results. *-s* is produced more for Non-Rhymes than for Rhymes, as emphasized by Clahsen (1999*a*); however, *-en* also shows the same directional preference, and at a much higher frequency.

Our rating results diverge from production results in some ways — for example, *-en* is favored instead of *-e* — and are consistent in others: both *-s* and *-en* are rated higher for Non-Rhymes compared to Rhymes. The low ratings for *-s* conflict with M95's findings, and suggest that presentation in sentence context is an important methodological difference from presentation in isolation. For example, family surnames obligatorily take *-s* in German, so

Plural	% All	Neut	M95 R	1 Syll
/-(e)n/	**37.3**	3.2	13.9	14.0
/-e/	34.4	**51.9**	**72.6**	**66.5**
ø	19.2	21.5	0.5	1.4
/-er/	2.9	10.6	7.3	4.7
/-s/	4.0	7.7	3.1	12.5
other	2.1	5.1	2.6	.9
N	11,243	2,606	642	570

Table 4.2: Distribution (percentages) of plural class for 1) nouns overall, 2) only neuter nouns, 3) nouns rhyming with M95 stimuli, 4) one-syllable nouns from Unimorph German dataset (Kirov et al., 2016).

it's possible that exposure to surnames in the "name" context primed subjects in the M95 rating study to find *-s* more acceptable generally, across conditions.[5] In any case, our results demonstrate task effects: although *-e* is the most *produced* plural form, *-en* obtains the highest *ratings* from the same speakers.[6] We compare these results with the modeling study in Section 4.2.4, focusing on production data.

4.2.3 Study 1b: Encoder-Decoder inflection

The second component of our study trains an encoder-decoder (ED) model on the task of German plural inflection, following the method of Kirov and Cotterell (K&C). We then compare its predictions on the M95 stimuli to the behavior of participants in Study 1a.

Background Wug tests have also been used to evaluate how computational models generalize; see §2.3 for general review, and §3.4 for an overview of efforts to model the German plural system. Our study follows Corkery et al. (2019) in aggregating production probabilities over several model initializations to compare these results to speaker production data.

Method

Overview We model German number inflection using the sequence-to-sequence Encoder-Decoder architecture (Sutskever et al., 2014). This comprises a recurrent neural network (RNN) which reads in an input sequence and **encodes** it into a fixed-length vector repre-

---

[5] Hahn (1999) reanalyze the M95 ratings and find that *-s* is rated much higher for family surnames than other kinds of names within the "name" condition (e.g. first names), reflecting the strong link between this category and the *-s* plural class.

[6] Further analysis indicates that individual survey participants rated a plural form they did *not* produce as better than the form they *did* produce in fully one-third of cases.

sentation, and another RNN which incrementally **decodes** that representation into an output sequence. Following Kann and Schütze (2016), our decoder uses neural attention.

For our task of morphological transduction, the ED model takes character-level representations of German nouns in their singular form as inputs (e.g. $\langle m \rangle$ H U N D $\langle eos \rangle$), and learns to produce the noun's inflected plural form (e.g. H U N D E $\langle eos \rangle$). Each character sequence starts with $\langle m \rangle$, $\langle f \rangle$, or $\langle n \rangle$, to indicate grammatical gender. Unlike English, the phonological-orthographic mapping is straightforward in German, so we can use a written corpus for model training. We keep a held-out dev set for hyperparameter selection, and a held-out test set to asses the model's accuracy in generalizing to unseen German nouns. In addition, the 24 M95 nouns were used for comparison with speaker behavior. They were presented to the model as neuter gender, consistent with Study 1a.

Corpus We trained all models on the UniMorph v1 German data set[7] (Kirov et al., 2016; Sylak-Glassman et al., 2015), which provides the singular and plural forms of 11,243 nouns. Only nominative case forms were used. Grammatical gender was obtained by merging the Unimorph dataset with a more recent Wiktionary scrape containing this feature.[8] Table 4.2 gives the distribution of plural suffixes for the UniMorph corpus overall, and for three relevant subsets: nouns with neuter gender, monosyllabic nouns (like the M95 stimuli), and nouns which were phonologically similar to the M95 stimuli, i.e. shared a rhyme. The number of items in the train, dev, and test splits is shown (in parentheses) in Table 4.3.

Implementation Following K&C and Corkery et al. (2019), our model is implemented using OpenNMT with their reported hyperparameters (after Kann and Schütze, 2016): 2 LSTM encoder layers and 2 LSTM decoder layers, 300-dimensional character embeddings in the encoder, and 100-dimensional hidden layers in both encoder and decoder; Adadelta optimization for training with a batch size of 20 and inter-layer dropout rate of 0.3; and a beam size of 12 for decoding during evaluation.

Since Corkery et al. (2019) found the ED model to be highly sensitive to initialization, we trained multiple simulations with the same architecture, varying only the random seed.

---

[7] `https://github.com/unimorph/deu`

[8] `https://github.com/gambolputty/german-nouns/` To ensure our results were not limited by the small size of the UniMorph dataset, we also trained the model on this larger dataset, including about 65,000 nouns. As the outcome was consistent with our findings here, we report results from the smaller model.

Train	Dev	Test
99.9% (8694)	92.1% (1229)	88.8% (1320)

Table 4.3: Model accuracy (N) by UniMorph corpus split, averaged over 25 random initializations.

	Test			M95		
	Prec.	Rec.	F1	%R	%NR	ρ
/-(e)n/	.95	.95	.95	6.3	3.3	.28
/-e/	.86	.89	.87	**68.3**	**91.7**	.13
ø	.96	.91	.92	0	0	
/-er/	.83	.85	.84	21.7	2.7	.05
/-s/	.64	.56	.60	3.7	2.3	.33
other	.37	.48	.42	0	0	

Table 4.4: Model results by plural suffix for: (left) test set performance (averaged over plural seed); (right) production percentages for rhyme (R) and non-rhyme (NR) M95 stimuli, and correlation (Spearman's ρ) to speaker productions.

Reported results combine predictions from 25 separate random initializations. The one hyperparameter we tuned was early stopping. Best performance on the validation set was achieved at 10 epochs, which was sufficient to memorize the training data.

Results The model achieves 88.8% accuracy on the held-out test set (Table 4.3). It performs best on *-en*, the most frequent class (Table 4.4). Unsurprisingly, the worst performance appears on the 'other' category, which comprises the long tail of idiosyncratic forms which must be memorized (e.g. Latinate plurals *Abstraktum* → *Abstrakta* or other borrowings *Zaddik* → *Zaddikim*). In keeping with the findings of Hahn and Nakisa (2000), *-s* is the plural suffix with the worst generalization performance; this cannot be attributed to low frequency alone (c.f. Table 4.2), as the model does much better on the similarly rare suffix *-er* .

We use the M95 stimuli to compare model predictions to speaker data from Study 1a. The model shows an overwhelming preference for *-e* on these words (Table 4.4); roughly 80% of its productions are *-e*, relative to 45% of speaker productions (Figure 4.1). In contrast, the model rarely predicts *-en*, which speakers use 30% of the time. The model's treatment of Rhymes and Non-Rhymes is even farther off the mark: where speakers use *-en* and *-s* *more* for Non-Rhymes relative to Rhymes, the ED model uses them *less*, producing *-e* for over 90% of Non-Rhymes. Following K&C and Corkery et al. (2019), we calculate the Spearman rank correlation coefficient (Spearman's ρ) between model and speaker production probabilities

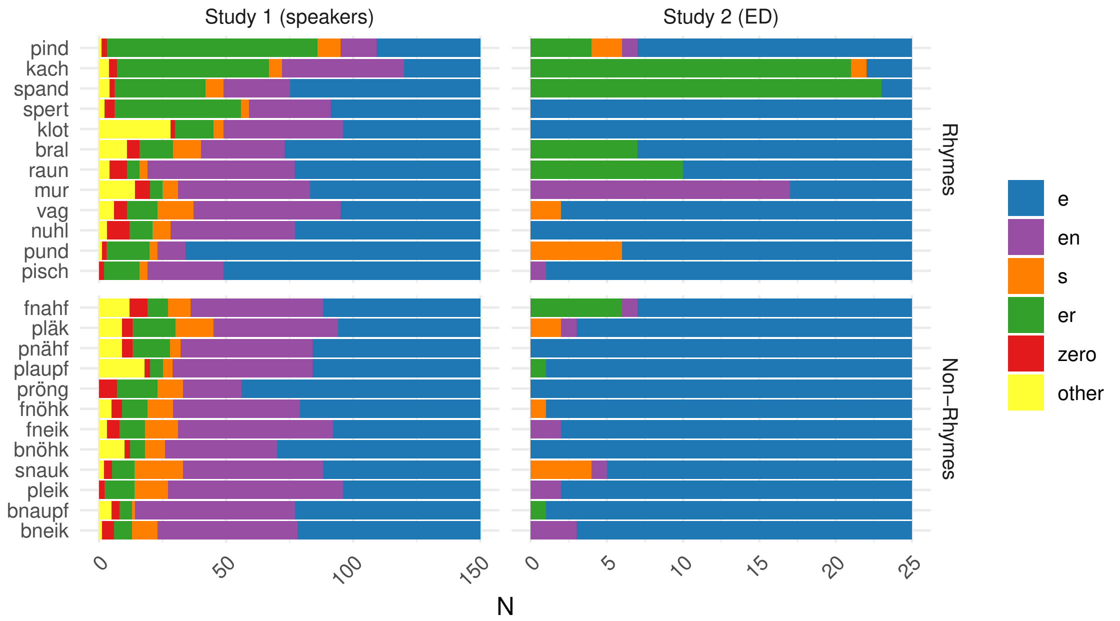


Figure 4.1: Plural class productions by item.

within inflectional categories rather than across categories.[9] This means that, for each potential plural suffix, we compare speaker and model productions for that suffix on each individual M95 word. Table 4.4 reports the correlation for each suffix. None show a statistically significant difference from the null hypothesis of no correlation.

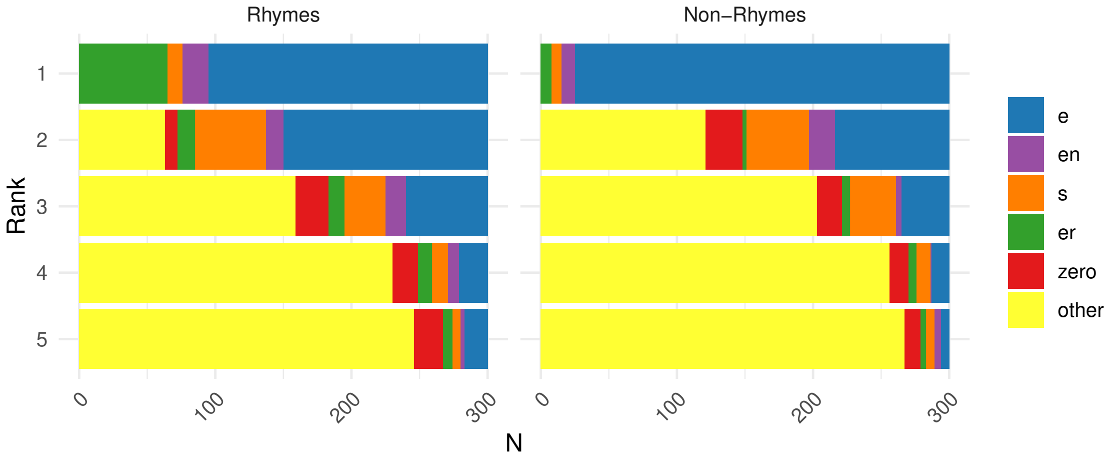


Figure 4.2: Distribution of plural classes by rank in ED model output.

Figure 4.2 shows the distribution of plural classes in the top 5 most likely forms predicted by the model for each M95 word. While all of the model's top-ranked predictions are well-formed

---

[9] For the English analyses in the prior works, this means calculating separate correlations for regular and irregular forms.

outputs in the sense that they conform to one of the main German plural classes, the lower-ranked predictions are rapidly dominated by "other" forms which do not cohere to standard plural production. An example from one model instance: the Rhyme input *Spert* had as its top five predictions *Sperte, Spelte, Spente, Sperten,* and *Fspern*; the Non-Rhyme input *Bneik* had *Bneiken, Bneiks, Bneikke, Bneikz,* and *Bneikme*. Corkery et al. (2019) observed instability in the ranking of irregular forms in ED models trained on the English past tense; however, English irregular forms are very diverse, which makes it difficult to draw broad conclusions about the plausibility of lower-ranked forms in the model's output. In contrast, the five main plural suffixes for German cover 98% of the nouns in the UniMorph dataset, and 95% of speaker productions on M95 stimuli in Study 1. The predominance of ill-formed plurals in lower-ranked predictions[10] suggests ED model scores may not be cognitively plausible analogues to speaker behavior; if they were, we would expect forms with standard plural inflections to receive consistently high rankings.

4.2.4 Discussion

The current study asks whether modern Encoder-Decoder neural models learn the full set of correct generalizations — that is, human-like behavior — with respect to German number inflection, which requires the learner to generalize non-majority inflectional classes. The short answer is no: our model learns *part* of that set. In particular, it correctly identifies -*e* as the 'best' plural class for this context. -*e* is the most frequent class in the training data for similar inputs (neuter gender, monosyllabic, phonologically close to M95; c.f. Table 4.2), and it is also the plural suffix most frequently produced by speakers (Table 4.1). Like all plural classes, -*e* does not characterize a majority of German nouns overall (Table 3.1), so the model has technically learned to generalize a minority class in its appropriate context. Nonetheless, it does not reproduce the behavior of survey participants in response to the same stimuli, which shows a more variable distribution over plural classes and different generalization patterns for Non-Rhymes relative to Rhymes.

This outcome is not surprising when one considers that the model is trained to produce one correct form rather than a distribution over plausible forms; however, this is exactly the task

---

[10] Interestingly, while less frequent classes such as -*s* and ø appear more often in the model's lower-ranked outputs, the class -*en* is almost never predicted — despite being the second most frequent class in speaker data productions.

faced by human language learners as well. All the models of morphology discussed here assume that exposure to correct forms alone should suffice for learning speaker-like behavior. Corkery et al. (2019, 3872, fn. 4) note that training on single target forms produces highly skewed ED model scores, with a great deal of probability mass on the top-ranked form and instability in lower rankings, but that training on a distribution would not be a cognitively plausible alternative. However, it could be the case that German speakers do regularly encounter variable realizations of plural forms. Köpcke observes that German plural inflection shows regional variation, for example northern speakers using *-s* (*die Mädels* 'girls') where southern dialects prefer *-en* (*die Mädeln*). Incorporating dialect-informed variability into training might be one way to encourage neural models toward speaker-like generalization.[11]

Parallel issues arise for model evaluation: how should we evaluate models of production when the target output is a distribution? On simplified versions of the task, such as classification (Hahn and Nakisa, 2000), the output distribution is constrained within a space of plausible forms, but sequence-to-sequence models deal with the open-ended domain of all possible strings. For encoder-decoders, the likelihood scores produced during beam-search decoding offer an intuitive option, and K&C use these scores to evaluate their model with respect to Albright and Hayes' wug data; however, Corkery et al. (2019) demonstrate that these model scores are not a suitable metric for that comparison. Our results provide further evidence that lower-ranked ED predictions do not reflect cognitively plausible distributions: they contain many ill-formed outputs, and omit inflectional classes such as *-en*, which is prevalent in speaker productions. An alternative to model scores is to treat each randomly initialized instance of a model as an individual, and compare aggregate productions with speaker data (Goebel and Indefrey, 2000; Corkery et al., 2019). For our experiments, this did not produce the distribution observed in the speaker data. The discrepancy between speaker production and rating preferences poses another challenge, as it's not clear how the ED model might represent these different task modalities.

Beside variability, the other key discrepancy between speaker and ED behavior is the treatment of Non-Rhyme words. If German has a default plural class, it should be realized more

---

[11] Like previous studies on these stimuli, our Study 1 did not collect data on speakers' dialect background; we are addressing this issue in follow-up research. We note that Study 1 began with an onboarding task prompting speakers to inflect existing nouns in Modern High German, which hopefully primed use of the standard variety for the following tasks.

often on these phonologically atypical stimuli than the more familiar Rhyme words. Speakers in Study 1 use *-s* and *-en* more for Non-Rhymes than for Rhymes. These results are consistent with earlier studies: M95 found that *-s* was the only plural form to receive higher average ratings for Non-Rhymes compared to Rhymes, and Z&L found that speakers produced both *-en* and *-s* more often for Non-Rhymes. In contrast, the ED model appears to treat *-e* as a default, producing *-e* inflections for under 70% of Rhymes but over 90% of Non-Rhyme inputs. This asymmetry suggests that the model has not induced the full set of correct generalizations for German plural inflection — it has not recognized which plural classes are more productive for phonologically atypical nouns. In fact, the model's preference for *-e*, the most frequent (if non-majority) suffix, is the behavior anticipated by M95: "frequency in the input to a pattern associator causes a greater tendency to generalize" (1995, 215). It seems that the productivity of less frequent inflectional classes continues to challenge neural models and limit their cognitive application.

4.2.5 Conclusions

German number inflection has been claimed to have distributional properties which make it difficult for neural networks to model. Our experimental speaker data does not necessarily support all of these claims; in particular, *-s* does not appear to be the only plural suffix which speakers treat as a 'default' for phonologically unfamiliar words, as the more frequent marker *-en* shows similar trends. Nonetheless, the German plural system continues to challenge ED architectures. Our neural model struggles to accurately predict the distribution of *-s* for existing German nouns. On novel nouns, it generalizes the contextually most frequent plural marker *-e*; its predictions are less variable than speaker productions, and show different patterns of response to words which are phonologically typical (Rhymes) as opposed to atypical (Non-Rhymes). Regardless of the minority-default question, it seems that ED models do not necessarily function as good cognitive approximations for inflectional systems like German number, in which no class holds the majority.

4.3 Additional Results

The following section presents additional modeling results from later experiments, and compares them to the findings reported by McCurdy, Goldwater and Lopez (2020). The modeling results (§4.3.1) motivate further analysis of the speaker data, also reported here (§4.3.2). This process aligns with the perspective advanced by Baroni (2022): computational models can be viewed as *algorithmic linguistic theories* which generate testable behavioral predictions.

The main finding presented here is that the recurrent neural encoder-decoder model under consideration (ED; Kann and Schütze, 2016; Kirov and Cotterell, 2018) *can* learn to generalize *-s* more to the phonologically atypical Non-Rhyme stimuli developed by Marcus et al. (1995), provided it is trained on wordforms alone — that is, *without* exposure to grammatical gender in the input. Furthermore, the model's generalization patterns are unexpectedly informative with respect to speaker behavior. The ED does not use *-s* for all of the phonologically atypical noun inputs, but only for a subset — specifically, nouns which end in *-k*. Modeling experiments with the symbolic ATP model proposed by Belth et al. (2021) show the same pattern. Further analysis of the speaker data reveals that nouns ending in *-k* likely drive the effect in *-s* production on these stimuli for speakers as well.

In other words, the findings presented here directly contradict the claim which concludes the introduction of McCurdy, Goldwater and Lopez (2020): "The model may correctly identify the most frequent class, but fails to learn the conditions under which minority classes are productive for speakers." It turns out that the ED can both correctly identify the most frequent class, and — under a certain training regime — correctly identify the conditions under which speakers generalize the minority class *-s*. We discuss the implications of this finding in §4.4.

4.3.1 Modeling Experiments and Results

Motivation

It is not immediately obvious why one would remove grammatical gender from the nouns provided as input to the ED. German speakers are undoubtedly exposed to the articles which mark noun gender, so removing indicators of gender creates an apparent asymmetry between the training environment of the neural model and that of human speakers.

The idea to train an ED model without grammatical gender was suggested by an anonymous reviewer for a gender-focused submission to AMLaP (McCurdy, Lopez and Goldwater, 2020*b*). That experiment is reported in Chapter 5. In the course of that analysis, we discovered that the ED trained on wordform alone showed a higher correlation to speaker productions than the ED trained on wordform and grammatical gender. This finding motivated us to analyze *-s* production for wordform-only EDs, leading to the results reported here.

We additionally train and evaluate the symbolic ATP model for two reasons. Firstly, Belth et al. (2021) analyze the speaker data released by McCurdy, Goldwater and Lopez (2020), and report that the ATP shows a higher correlation to speaker productions than the ED discussed in our original paper. This suggests that, like the ED trained without gender, the ATP shows a better statistical approximation to speaker behavior. Their model, however, was trained on a smaller amount of data drawn from a different source (namely, a noun lexicon extracted from child-directed adult speech in CHILDES; MacWhinney and Snow, 1985) — we can't be sure that this advantage holds when using the same training data as the ED. Secondly, the ATP proved capable of learning an *-s* default rule for German plural inflection, at least when trained on the acquisition data extracted by Belth et al. Under the theory advanced by Marcus et al. (1995), this capacity should substantially contribute to speaker-like generalization. By training ATP models on the same data as our EDs, we can see whether it also learns an *-s* default rule in this setting — and if so, whether application of the *-s* default rule leads to more speaker-like predictions compared to the ED.

ATP Model

The "Abduction of Tolerable Productivity" (ATP) model proposed by Belth et al. (2021) is a rule-based symbolic learner. In keeping with the classical linguistic tradition of ordered rules (e.g. Chomsky and Halle, 1968), each rule comprises an input condition (expressed on the Left-Hand Side: LHS) and associated structural change (on the Right Hand Side: RHS). As a running example, we will consider a rule which assigns the *-(e)n* plural to feminine German nouns. This rule might look as follows: $[+\text{F}] \rightarrow$ *-(e)n*.

Given an input lexicon of wordforms and features, where each word has a paired structural transformation (e.g. a plural inflection class as shown in our example), the ATP generates candidate rules based on the combinations of features and transformations observed in the data, and selects a winner based on the Tolerance Principle criterion described below. The words which match the input condition (the LHS) of the winning rule are then removed from consideration, and the process is repeated on the remaining data. For example, if the [+F] → *-(e)n* rule happened to be selected, the ATP would then remove all feminine nouns from consideration (including exceptions, i.e. feminine nouns which do *not* take *-(e)n*) and generate a new set of candidate rules based on the remaining data. This process recursively partitions the input lexicon, resulting in a decision tree of productive rules and stored exceptions.

The Tolerance Principle (Yang, 2016) provides a simple data-driven criterion for rule application. The rule threshold is $\frac{N}{\ln(N)}$, where N is the number of items (in our case, word types) which match the LHS input condition of a candidate rule. According to the Tolerance Principle, if the number of exceptions to a rule is less than $\frac{N}{\ln(N)}$, then the rule is productive. Consider our running example [+F] → *-(e)n*. Within a particular dataset or dataset partition, this rule is productive if the number of feminine nouns which take *-(e)n* (i.e. the nouns for which the rule is true) is greater than the Tolerance Principle criterion (i.e. $N - \frac{N}{\ln(N)}$, where N is the total number of feminine nouns). Yang (2016, §4.4) applies the Tolerance Principle to German plural inflection, using a noun lexicon drawn from child-directed speech. In his analysis, [+F] → *-(e)n* is the first rule to reach productivity under this criterion.

The ATP combines the Tolerance Principle criterion with a system to generate and select candidate rules in the form described above. The LHS of candidate rules are generated by exhaustive search through two possible types of input conditions: 1) orthographic sequences of 1-6 characters at the ending of the input wordform, or 2) morphological features provided as labels with the data. For example, the set of features under consideration for the noun *Reservierung* would comprise the set of stem endings *{g, ng, ung, rung, erung, ierung}* and the feminine gender feature *+F* if included in the input. The RHS of candidate rules comprise all transformations observed with those features in the data; in the case of *Reservierung*, this means each LHS feature generates a candidate rule mapping to the *-en* plural. Candidate rules can be disjunctively combined when the all input conditions are orthographic features which independently meet the Tolerance Principle criterion, e.g. [ung‖heit‖keit] → *-en*. If multiple candidate rules pass the Tolerance Principle criterion, the ATP selects the rule which has the fewest exceptions. If no candidate rules are productive by this criterion and N is small, the ATP falls back on majority rule. If N is big on the LHS, or there is no majority output class on the RHS, all the words which match the input condition are stored as exceptions under a "failed" node. By design, the ATP is effectively deterministic for a given dataset: it will always learn the same decision tree over a particular set of input words.

An ATP model which has learned a particular decision tree can classify words not observed in its vocabulary. The inference procedure starts at the first rule in the decision tree. If the new word matches the input condition of the first rule, and does not appear on the list of memorized exceptions to that rule, then the ATP predicts the output transformation associated with that rule. If the word does not meet the input condition, then it is evaluated at the next node in the tree. If the word reaches a "failed" node where no productive rule has been identified, the ATP assigns the output classification of the most similar word stored at that node, based on string edit distance.[12]

Method

To train the wordform-only neural model, we use the same method as McCurdy, Goldwater and Lopez (2020), with one key difference: grammatical gender is not included in the input to the model during training or evaluation. The ED model takes character sequences of German nouns in their singular form as inputs (e.g. H U N D), and learns to produce the noun's inflected plural form (e.g. H U N D E). We use the same corpus (UniMorph) with the same training, validation, and test splits. As before, we train 25 unique model instances with separate random seeds.

The ATP requires a different training regime, as it is effectively deterministic for a given dataset and requires no hyperparameter tuning. Since the validation set is not needed for hyperparameters, we combine it with the training set, yielding a lexicon of 9923 nouns. We then draw 25 independent samples of 8694 nouns, the same size as the training set used for the ED. 25 separate instances of the ATP are trained on these samples. For comparison with the RNN, we train the ATP under two different regimes: one with grammatical gender and wordform, and one on wordform alone. In this chapter, we consider the predictions of the wordform-only model, but both versions of the ATP are evaluated in Chapter 7.

Following McCurdy, Goldwater and Lopez (2020), both models are evaluated on the novel noun stimuli developed by Marcus et al. (1995).

---

[12] This procedure means that the ATP might qualify as a dual-mechanism model, as discussed in §3.2.1.

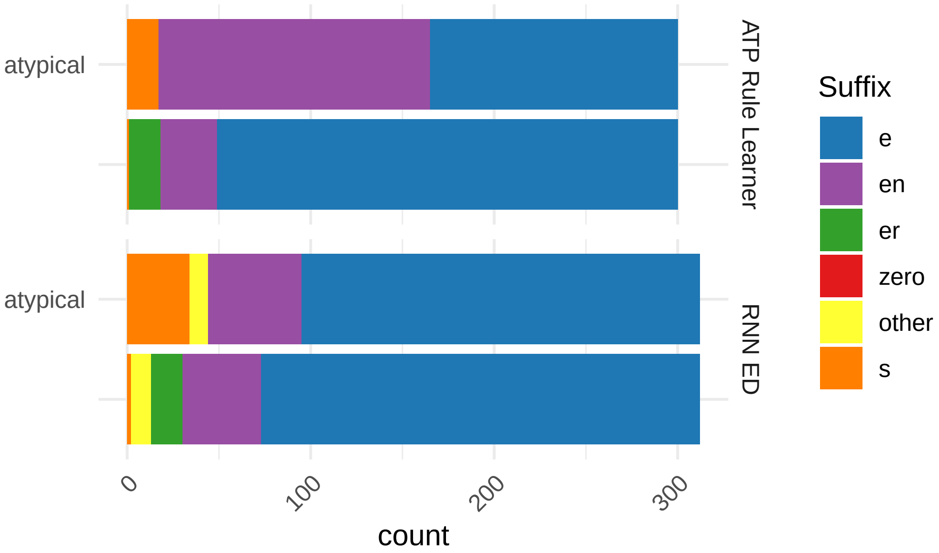


Figure 4.3: Distribution of predicted plural class for Marcus et al. stimuli. Both the neural RNN Encoder-Decoder (ED) model and the symbolic ATP model predict *-s* more frequently for phonologically atypical nouns ("Non-Rhymes") compared to typical stimuli ("Rhymes").

Results

We find that, when trained on wordforms only, both the ED and the ATP model produce *-s* more often for phonologically atypical stimuli (Fig. 4.3). To assess whether this pattern is statistically reliable, we fit a separate binomial generalized linear mixed effect model to the ED and ATP predictions using the LME4 package (Bates et al., 2015) in R (R Core Team, 2023), with random intercepts for model instance (seed) and item (input wordform), and phonological typicality (rhyme vs. non-rhyme) as the fixed effect. We find that the ED produces *-s* significantly more for phonologically atypical nouns $(\beta = 3.53, std.err = 1.35; z = 2.6, Pr(> |z|) = 0.0091 **)$, but the effect is not significant for the ATP $(\beta = 2.68, std.err = 1.98; z = 1.35, Pr(> |z|) = 0.18)$.

The -k confound Fig. 4.4 plots the distribution of plural class predictions for each item in the evaluation stimuli. If we focus on *-s* predictions, a clear pattern emerges: both the ED and ATP predict *-s* almost exclusively for input words ending in *-k* (e.g. *Snauk, Pläk, Bneik*), even though no single item receives a majority of *-s* predictions from either model. All of the stimuli which end in *-k* happen to fall in the category of phonologically atypical nouns, which strongly suggests that this confound drives the typicality effect described above.

Statistical analysis further supports the appearance of a *-k* effect, rather than a phonological

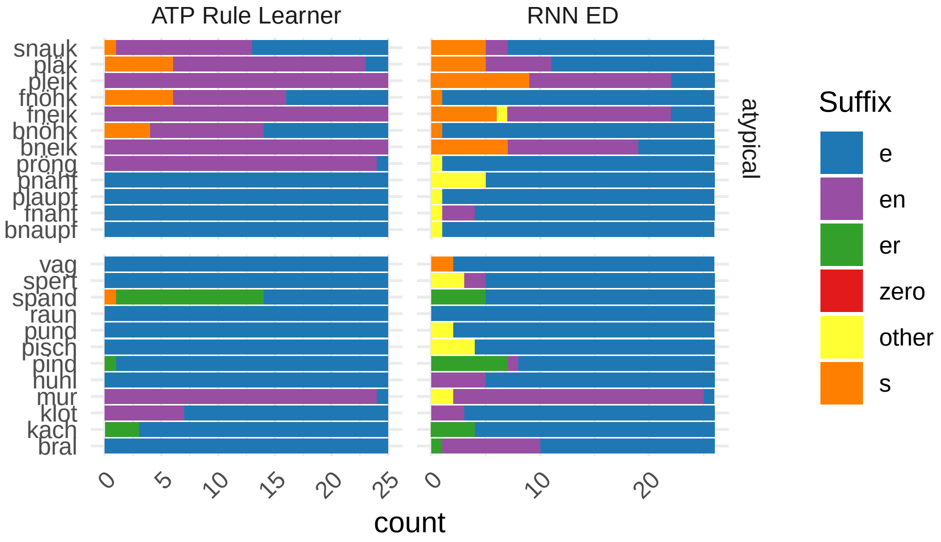


Figure 4.4: Distribution of predicted plural class for Marcus et al. stimuli by item. Both the neural RNN Encoder-Decoder (ED) model and the symbolic ATP model predict *-s* almost exclusively for nouns ending in *-k*, which all happen to be phonologically atypical nouns.

typicality effect. We fit another binomial mixed effect model to the prediction data with the same structure described above, but this time we use a binary variable for noun ending (*-k* vs. other) as the fixed effect rather than phonological typicality. This analysis finds significantly higher *-s* production for *-k*-ending nouns for both the ED ($\beta = 4.13, std.err = 0.87; z = 4.8, Pr(> |z|) = 0.000 * **$) and the ATP ($\beta = 4.2, std.err = 1.5; z = 2.8, Pr(> |z|) = 0.005 * *$). Unsurprisingly, an ANOVA comparison finds that *-k*-ending is a better statistical predictor than phonological typicality in both cases. The Bayesian Information Criterion (BIC) for models with phonological typicality is 237 for the ED predictions and 143.7 for the ATP; using *-k*-ending as the fixed effect instead yields lower BICs of 222.3 and 138.6 respectively, indicating a better model fit. Adding phonological typicality does not significantly improve the fit of either generalized linear model. These findings show that lemmas ending in *-k* are the clear statistical driver of *-s* assignment for both the neural ED and symbolic ATP.

The origin of the *-k* effect is apparent if we inspect the UniMorph German noun lexicon used to train both models. In UniMorph, approximately 15% of nouns ending in *-k* take the *-s* plural suffix. This is by far the highest proportion relative to all the other singular noun endings represented in our novel noun stimuli (Table 4.5).

k	t	n	f	l	d	g	r	h
15.1	4.4	3.0	2.7	2.4	1.4	1.3	1.3	0.7

Table 4.5: Percentage of singular nouns with a given ending which take an *-s* plural, calculated from UniMorph v1 Sylak-Glassman et al. (2015).

Does the ATP learn a minority default representation? One of the motivations for training and evaluating the ATP is its capacity to learn a default *-s* rule, as reported by Belth et al. (2021). Following their interpretation, we consider an ATP model to have a default rule if and only if it assigns a particular class when no input condition is met (the elsewhere condition; §2.1.3). We manually inspect the learned decision tree for each of the 25 trained ATP instances and find that none of them learn a default rule for any plural class, *-s* or otherwise. In fact, when trained on wordform alone,[13] *no* ATP instance learns a productive rule to generalize the *-s* plural, which implies that the *-s*-*k* association found in these results is driven by the ATP's analogical fallback mechanism.

4.3.2 Speaker Data Analysis

Motivation

The analysis in the preceding section demonstrates that a neural model of German plural inflection, when trained on the orthographic form of words alone, can capture the key speaker behavior found by both Marcus et al. (1995) and McCurdy, Goldwater and Lopez (2020): a statistically reliable increase in *-s* assignment for the phonologically atypical nouns in their stimuli. Furthermore, both neural and symbolic computational models predict that *-s* assignment should largely appear on those atypical stimuli which end in *-k*. The latter finding motivates our current investigation: is the *-s* typicality effect we observe in speaker behavior actually a *-k*-ending effect instead?

Given that speaker behavior is much more variable than both models' predictions, we might not expect to see such a clean divide. The difference between speaker and model behavior is obvious when one visually compares Fig. 4.4 to Fig. 1 of McCurdy, Goldwater and Lopez (2020). For every item in our stimuli, one or more speakers produce an *-s* plural form; compare this to the RNN and ATP, which never predict *-s* for most items. However, despite the increased

---

[13] When trained on wordform and grammatical gender, the ATP sometimes learns productive rules to generalize *-s*; however, they are never default rules. For example, one ATP instance formalizes the *-s*-*k* association as part of a rule generalizing *-s* to words ending in *-a, -i, -k, -o, -p, -u, -y, -on, -pot*.

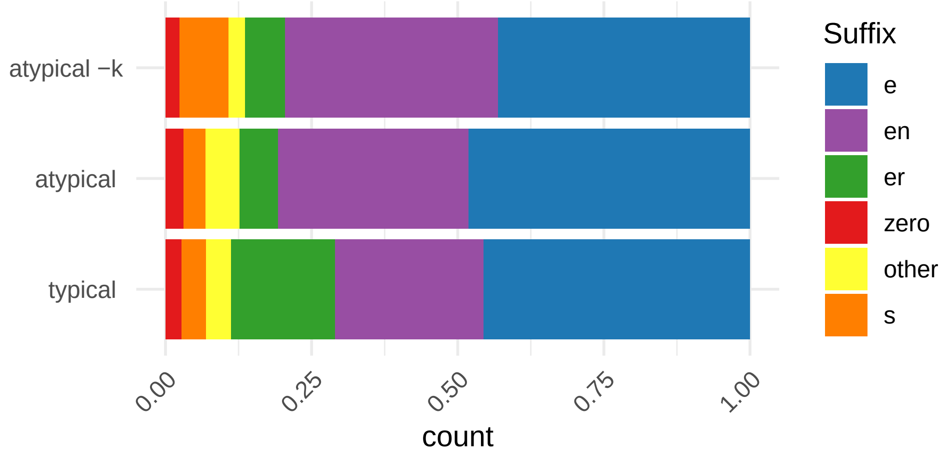


Figure 4.5: Speaker productions from McCurdy, Goldwater and Lopez (2020), with phonologically atypical nouns further subdivided by ending with -*k*. Speakers produce -*s* more frequently for the 7 atypical nouns ending in -*k*; for the other 5 atypical nouns, speakers produce -*s* at rates equivalent to the 12 typical nouns.

variability in speaker data, they do appear to produce -*s* more often for nouns ending in -*k*, as shown in Fig. 4.5.

Reanalysis of McCurdy, Goldwater and Lopez (2020)

We follow the same statistical procedure as above to evaluate whether phonological typicality or -*k*-ending provides a stronger account of speaker -*s* production. Using a binomial generalized linear mixed effects model with random intercepts for item and participant, we find the same phonological typicality effect reported by McCurdy, Goldwater and Lopez (2020): speakers produce -*s* significantly more for phonologically atypical words (i.e. Non-Rhymes; $\beta = 0.6, std.err = 0.3; z = 1.98, Pr(> |z|) = 0.048*; BIC = 1158.1$). However, similar to the model results discussed in §4.3.1, we find a larger effect and better statistical fit if we use ending in -*k* as the fixed effect predictor instead ($\beta = 1.09, std.err = 0.28; z = 3.94, Pr(> |z|) = 0.0001 * **; BIC = 1149.2$). If we include both -*k*-ending and phonological typicality as main effects in our model, it finds a significant effect for the former ($\beta = 1.2, std.err = 0.37; z = 3.19, Pr(> |z|) = 0.001 * *$), but not the latter($\beta = -0.15, std.err = 0.35; z = -0.42, Pr(> |z|) = 0.67$). Furthermore, an ANOVA comparison finds that the model with both typicality and -*k*-ending as predictors does not provide a significant improvement over the model with only -*k*-ending ($\chi^2 = 0.18; Pr(> \chi^2) = 0.67$).

These results indicate that speakers, like the models discussed above, tend to increase -*s*

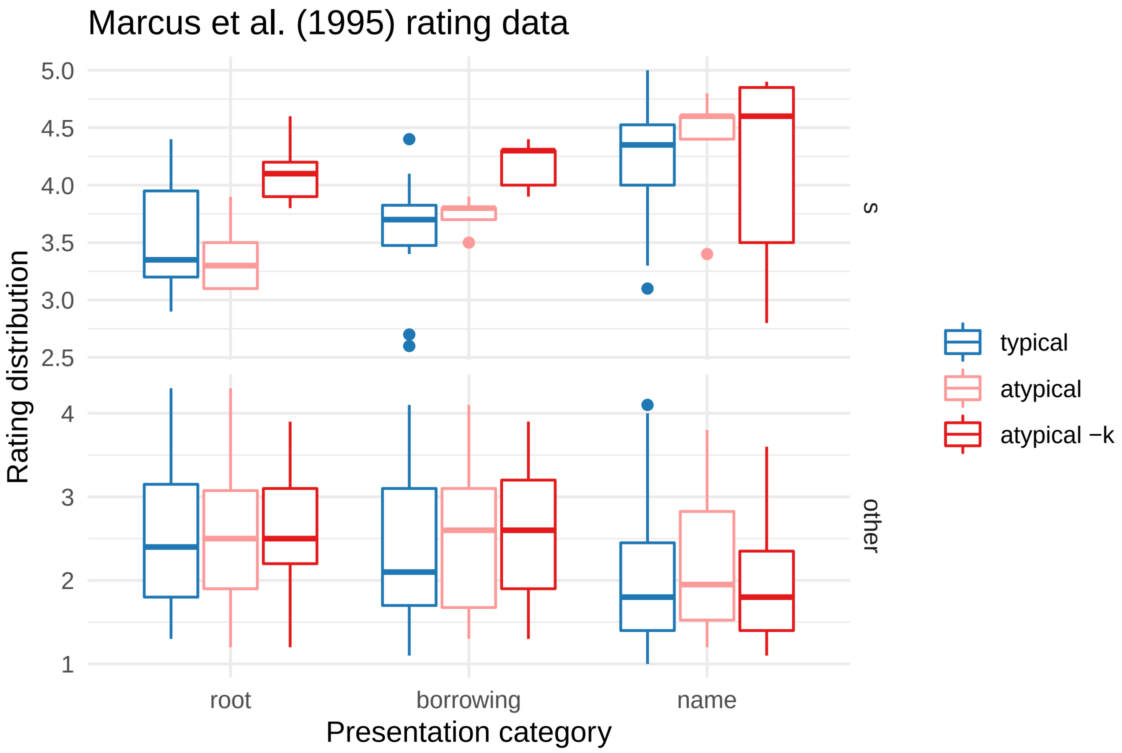


Figure 4.6: Distribution of average item ratings by phonological typicality and presentation category (data from Marcus et al., 1995, Appx. 3). For *-s* plural forms (top row) in the Root and Borrowing categories (upper left / upper middle), average ratings for atypical nouns ending in *-k* are significantly higher than both typical nouns and atypical nouns with other endings. This *-k*-ending asymmetry does not appear for other plural endings (lower row). There is also no asymmetry for *-s* plural forms in the Name category (upper right), where ratings are generally high because *-s* is the required plural class for German proper names.

production for lemmas ending in *-k*. Because *-k*-ending is confounded with phonological atypicality in the Marcus et al. stimuli, it appears that speakers produce *-s* more for phonologically atypical Non-Rhymes. However, phonological typicality has no predictive power independent from ending in *-k*; when this confound is directly included in the model, we find no statistical support for the hypothesis that speakers produce *-s* more for atypical Non-Rhymes.

Reanalysis of Marcus et al. (1995)

The item-level rating data published by Marcus et al. (1995, Appx. 3) lends further support to the hypothesis that *-s* preference is driven by *-k*-ending words rather than phonological typicality. As shown in Figure 4.6, participants gave higher average ratings to *-s* plural forms for nouns ending in *-k* in the Root and Borrowing categories, while all nouns that did not end in *-k* received similar *-s* plural ratings regardless of phonological typicality. This *-k*-ending asymmetry in *-s* plural rating is not apparent in the Name category, but its absence likely reflects a ceiling effect: *-s* is the grammatically required plural class for proper names in German, so nearly all *-s* plural forms received high ratings in this category.

To further evaluate the *-k*-ending hypothesis, we perform a statistical reanalysis of the rating data. Following the interpretation of Zaretsky and Lange (2016), we assume that the experimental presentation of McCurdy, Goldwater and Lopez (2020) is likely closest to Marcus et al.'s Root category: participants presented with bare nouns might consider them as German root words, but would not likely consider them borrowed nouns or proper names without contextual evidence. For this reason, we focus on rating data from the Root category. In their original ANOVA analysis of nouns presented as Roots, Marcus et al. found that *-s* plural ratings were higher for atypical Non-Rhymes compared to typical Rhymes: "the difference was significant by subjects, $F(1, 47) = 7.69, p < .01$, though not by items, $F(1, 22) = 5.47, p = .16$" (1995, p. 237). Without access to fine-grained data, we cannot perform the mixed-effects analysis as in the previous section. We can, however, use the average ratings to perform item-level ANOVA analysis and compare the results.

Our statistical reanalysis supports the *-k*-ending hypothesis. We first run an ANOVA by items to reproduce the result reported above. For *-s* plural forms in the Root category, we do not find significant effect of phonological typicality on average rating $(F(1, 22) = 2.20, p = .15)$. This is broadly the same result reported by Marcus et al., despite a minor discrepancy between the two calculated F values. If, however, we instead consider the effect of ending in *-k*, the ANOVA reveals a significant effect on *-s* plural ratings $(F = 11.82, p = .002)$. We find further evidence for the importance of *-k*-ending items through model comparison. The first ANOVA model, with only phonological typicality as a predictor, sees a significant improvement in model fit by adding *-k*-ending as an independent predictive factor $(F = 8.83, p = 0.007)$. The converse, however, is not true: if we start with *-k*-ending as the only predictor, we see no significant gain in model fit by adding phonological typicality $(F = 0.34, p = .57)$. We find a consistent pattern in both the rating data collected Marcus et al. (1995) and the production data collected by McCurdy, Goldwater and Lopez (2020): while there is some statistical evidence that German speakers prefer *-s* plurals for phonologically atypical Non-Rhyme stimuli, we have stronger evidence that speakers prefer *-s* for nouns ending in *-k*, and this confound most likely drives the typicality effect in both studies.

4.4 Discussion

The additional results presented in §4.3 lead us to critically revise the conclusions of McCurdy, Goldwater and Lopez (2020) with respect to both speaker and model behavior.

In terms of speaker behavior, we find no support for the hypothesis that speakers prefer to generalize *-s* to nouns which are phonologically atypical. Instead, our statistical analysis indicates that speakers have a slightly higher tendency to generalize *-s* to nouns which end in the letter *-k*, which is confounded with phonological atypicality in the stimuli developed by Marcus et al. (1995). The *-k* confound means that these stimuli are not suitable to evaluate the effects of phonological typicality. In principle, it is still possible that German speakers may show some systematic behavior in generalizing plural inflection to phonologically atypical novel nouns, in a way that differs from generalization to phonologically typical nouns; however, the behavioral evidence presented by Marcus et al. (1995) and McCurdy, Goldwater and Lopez (2020) does not provide any statistical support for this hypothesis. In fact, having accounted for the lexical confound driving the apparent phonological typicality effect, we are left with a highly variable distribution of inflection class assignments — this variation appears to be the most stable and noteworthy aspect of how German speakers generalize plural classes to these stimuli. We discuss this and its implications for the minority default hypothesis below.

On the computational modeling side, we find both neural and symbolic models trained on orthographic wordform alone can achieve speaker-like generalization of the *-s* plural class: like German speakers, they prefer to assign the *-s* plural class slightly more frequently to novel nouns ending in *-k*. Both models still differ from speakers by showing less variation in plural class generalization broadly, but they capture the key conditions informing *-s* generalization. This raises an obvious question: why did McCurdy, Goldwater and Lopez (2020) find such different results when evaluating an ED trained on wordform and grammatical gender? We consider this question below, and further explore the effects of grammatical gender on ED generalization in Chapter 5.

4.4.1 No evidence for minority default in speaker plural generalization.

Our statistical analysis of speaker behavior in §4.3.2 indicates that ending in *-k*, rather than phonological atypicality, drives the *-s* effect found by Marcus et al. (1995) and McCurdy, Goldwater and Lopez (2020). What are the implications for the minority default hypothesis?

Marcus et al. (1995) present two different types of evidence for the minority default analysis of *-s* in German: linguistic analysis of the distribution of *-s* in the German noun lexicon (§3.2.1), and results from their behavioral experiments (§3.3.2). Of these two, the status of the former evidence is not under dispute. It is a fact about the German language that, compared other plural classes, the *-s* plural suffix appears in a more diverse and heterogeneous set of phonological and grammatical environments; this is apparent in the descriptive lexical statistics reported in §3.1, for instance Table 3.3 and Figures 3.1 and 3.2. The behavioral experiment is designed to yoke these two broad categories of evidence together by demonstrating that speakers generalize the *-s* form to novel words in ways that reflect this heterogeneous lexical distribution.

Marcus et al.'s experiment targets three different aspects of the "elsewhere distribution" of *-s* — its application to proper names, to lexical borrowings from other languages, and to phonologically atypical nouns which are dissimilar to the existing noun lexicon (see examples in Table 3.7). Of these three categories, the latter effect of phonological typicality has generally been taken to provide the strongest evidence in favor of the minority default hypothesis. Proper names generally take the *-s* plural class in German, but it is not obvious that this reflects "default" circumstances as opposed to special treatment of proper names, which occurs in many languages (e.g. Indefrey, 1999; Stemberger, 2006). Borrowings are a somewhat more complicated category. It is difficult to conceptually distinguish borrowed and phonologically atypical words: words from other languages often have phonological properties which are not found in the language "receiving" the new term, and this phonological atypicality is often invoked to explain special treatment of borrowed words. On the other hand, Marcus et al. found that their participants assigned higher ratings to *-s* plural forms of the same words when they were presented as borrowings rather than roots, and this difference was statistically significant. This result suggests an effect of lexical borrowing which is independent of phonology. Nonetheless, phonological atypicality arguably accounts for a large portion of the borrowing effect. Consistent with this premise, we observe that participants in the Marcus et al. study gave higher ratings on average to *all* plural forms of atypical nouns in the "Borrowing" category, and this effect nearly reaches statistical significance ($F(1, 154) = 3.84, p = 0.052$). It seems that phonologically atypical words may be more plausibly considered as borrowings regardless of their plural inflection class. Given the disputed status of proper names and strong interaction between phonology and borrowing, the phonological typicality effect becomes Marcus et al.'s core experimental evidence showing that speakers generalize *-s* under default conditions.

The Brain and Behavioral Sciences review by Clahsen (1999*b*) further confirms that phonological atypicality is central to the minority default analysis of *-s* generalization.[14] In §4.2.1, he points to the phonological typicality effect found by Marcus et al. to support the claim that "*-s* plurals are applied elsewhere, even to nouns that are not similar to any existing German word" (1999*b*, 997). In §5.1.2, he cites a rating experiment using the same stimuli with German-speaking children, which found they also preferred *-s* plurals for phonologically atypical Non-Rhymes as well as proper names (Bartke et al., 1995). In his author response, Clahsen again highlights the phonological typicality effect found by Marcus et al. as evidence that "adults do indeed generalize -s plurals to nonrhyming real words and to foreign words in German" (1999*a*, 1049). And in §4.6, he critiques the neural models developed by Hahn and Nakisa (2000) and Goebel and Indefrey (2000) for their failure to generalize *-s* under default circumstances. In particular, Goebel and Indefrey's recurrent model learned to apply *-s* to nouns ending in *-o*, which are phonologically similar to existing *-s* plural words; however, it "failed to generalize *-s* under no-similarity conditions," i.e. to phonologically atypical Non-Rhymes, and thus "does not capture the generalization properties of *-s* plurals" (1999*b*, 1005).[15] The message is clear: *-s* generalization based on phonological similarity, i.e. positive association with existing words, is not consistent with the minority default hypothesis — speaker-like *-s* generalization must reflect only dissimilarity, i.e. phonological atypicality or negative association with existing words, to constitute default elsewhere application.

Our reanalysis of speaker data in §4.3.2 indicates that, due to confounded stimuli, neither Marcus et al. (1995) nor McCurdy, Goldwater and Lopez (2020) provide behavioral evidence for the minority default hypothesis of *-s* generalization expressed by Clahsen (1999*b*). Instead, speaker data from both studies points toward a numerically small but robust positive association between *-s* generalization and nouns ending in *-k*, an association which reflects lexical statistics

---

[14] Note that Clahsen (1999*b*) discusses additional experimental evidence for the exceptional status of the German *-s* plural based on existing words, including behavioral evidence from lexical decision tasks in different modalities, and neuro-imaging experiments in which speakers read sentences with incorrectly inflected noun forms. Our discussion centers on the role of *-s* in speaker *generalization* to novel words, so we restrict our consideration to experiments which share this focus.

[15] As discussed in §3.4, Goebel and Indefrey's model primarily learned a gender-conditioned default generalization strategy.

(c.f. Table 4.5) and can be learned by both neural and symbolic models of German plural inflection (c.f. §4.3.1). These findings are broadly consistent with the Zaretsky and Lange's large-scale production study using the same stimuli. We could not evaluate the alternative -*k*-ending hypothesis on their study as they did not release sufficiently fine-grained data; however, their item-level analysis shows that the word-final phoneme of the singular noun has a significant effect on speaker production of -*s* plural forms, an effect consistent with our interpretation (2016, 167, Table 3). We are left with no robust behavioral evidence in favor of the minority default hypothesis that speakers generalize -*s* based on phonological atypicality.

Limitations and future directions The analysis discussed above shows that -*k*-ending, rather than phonological typicality, accounts for the difference in -*s* generalization on Marcus et al.'s stimuli; once -*k*-ending is controlled for, we see no systematic distinction in how speakers treat Rhyme and Non-Rhyme items. This evidence is compatible with the null hypothesis, i.e. no effect of phonological typicality, but it cannot *prove* this hypothesis. Demonstrating this null hypothesis would require testing with a broader range of stimuli carefully controlled against such confounds to isolate the effect of phonological typicality, for instance systematically varying phonological typicality for a range of stimuli ending in -*k*. Nonetheless, §4.3.2 refutes the key piece of behavioral evidence for Marcus et al.'s minority default theory of the German plural system. It shows that a statistical frequency association in the lexicon (between -*k* and -*s*) better explains -*s* generalization than than the atypical elsewhere distribution, thus supporting the predictability and type frequency criteria over the rule generation criterion for morphological regularity (§2.1.1). As the burden of proof lies with proponents of the minority default hypothesis, I consider this finding adequate for the purposes of this dissertation.

4.4.2 Models are overconfident relative to speakers.

Marcus et al. (1995) predicted that neural network models of morphological inflection would fail to achieve speaker-like generalization of the German -*s* plural because, under the elsewhere condition (§2.1.3), these models prefer to generalize frequent classes instead.

> The type frequency hypothesis exploits a correlation: the inflectional form that serves as the default in generalization is also the majority form in English. The hypothesis assumes a causal relation: frequency in the input to a pattern associator

> causes a greater tendency to generalize. (Marcus et al., 1995, 215)
>
> The predictions of the theory that all inflection is computed in a single pattern associator are different, because pattern associators neither easily generalize low-frequency suffixes, nor unite the different default circumstances (phonological and derivational) as defaults. [...]
>
> These models predict that *-s* should be eschewed across the board: driven only by phonological similarity, the models should always prefer the more common *-e*, *-(e)n*, and *-er* plural forms to *-s*, even for Non-Rhymes, since there is no reason that very rare *-s* would scoop up the words that have lower similarities to existing irregulars and no similarity to existing regulars. [...]
>
> Given the rarity of *-s* plurals both typewise and tokenwise, though, it is unclear whether it would learn that *-s* is also more applicable to unusual roots. (Marcus et al., 1995, 233)

We can separate this prediction into two key claims. The first claim is that pattern associators — i.e. neural network models — should not learn to assign the *-s* plural to Marcus et al.'s phonologically atypical Non-Rhyme stimuli. This claim about *-s* is empirically false: as shown by the results reported in §4.3.1, a neural ED trained without grammatical gender can exploit the *-k* stimulus confound to produce more *-s* plurals for Non-Rhymes. It does not appear to treat *-s* as the minority default class; however, this is unproblematic given the analysis in §4.3.2, which indicates that German speakers also appear to generalize the *-s* plural based on the *-k* confound, rather than treating it as a minority default. Nonetheless, it is true that the ED does not *reliably* show speaker-like *-s* generalization: McCurdy, Goldwater and Lopez (2020) found that an ED trained with grammatical gender did not increase *-s* plurals for Non-Rhyme *-k*-ending stimuli.

The second claim is broader, and independent of the first: neural network models should tend to generalize higher-frequency classes based on phonological similarity. Based on the results reported in McCurdy, Goldwater and Lopez (2020) and §4.3.1, this perspective appears generally correct, and accounts for some key properties of the ED's plural generalization which differentiates its behavior from the speakers in our experiment. In particular, model predictions appear overconfident relative to speaker generalization behavior.

Grammatical gender affects neural models' *-s* generalization.

Why does a neural ED trained on wordform alone learn to generalize *-s* to nouns ending in *-k* (§4.3.1), while an ED trained with wordform and grammatical gender does not (§4.2; McCurdy, Goldwater and Lopez, 2020)? While it is difficult to say precisely, we know that grammatical gender has a strong statistical association with inflection class in the German lexicon (Williams et al., 2020), and we have evidence that neural network models are particularly sensitive to this association (e.g. Goebel and Indefrey, 2000; McCurdy and Serbetçi, 2017; Dankers et al., 2021). From this, we can speculate that the ED may attend to grammatical gender at the expense of more subtle phonological cues, such as the relatively slight association between nouns ending in *-k* and the *-s* plural class. This might be true even though all the evaluation stimuli were presented to the model with the same grammatical gender, i.e. neuter (*Das*); the mere presence of gender in the training data may suppress the model's sensitivity to other cues.

Does grammatical gender affect speaker generalization behavior in a similar way? McCurdy, Goldwater and Lopez (2020) did not vary grammatical gender in presenting stimuli to speakers, so that data cannot address this question. Marcus et al. (1995) did not evaluate gender effects in their experiment, reporting only that grammatical gender did not show any interactions with other factors of interest. Zaretsky and Lange (2016) used the same stimuli in their production experiment, and found that grammatical gender yielded the most robust statistical influence on plural class assignments; however, this did not appear to suppress speakers' tendency to generalize *-s* to phonologically atypical Non-Rhymes. Chapters 5 and 6 present further behavioral experiments to investigate the effect of gender on speaker plural production.

Models, but not speakers, use phonological similarity to regularize.

Although it seems that computational models can capture the conditions under which speakers generalize the *-s* class, if we visually compare their item-level productions in Fig. 4.4 to the speaker data in Fig. 1 of McCurdy, Goldwater and Lopez (2020), we see that an interesting divergence remains. At the population level, German speakers in our behavioral experiment tend to produce a diverse and relatively consistent distribution of plural forms for each item in the set of stimuli. By contrast, both the neural and symbolic model produce far more variability

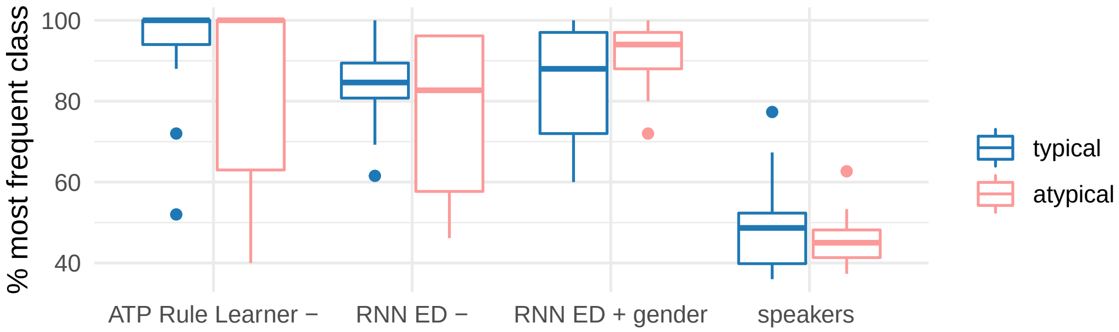


Figure 4.7: Percentage of productions per item for most frequent class for that item. "-" indicates a model trained on wordform alone, while "+ gender" indicates a model trained on wordform and grammatical gender. For any stimulus item, roughly 80-90% of the neural RNN ED and symbolic ATP model instances agree on the plural class for that item, while on average 45-50% of speakers assign the same class to one item.

between items, and far less variability *within* items. -*s* productions are a good example of this. Both the ED and the ATP predict -*s* plural forms almost exclusively for items ending in -*k*. The ED has an -*s* prediction rate of 19% for -*k*-words compared to .5% for all other words, and the respective ATP -*s* prediction rates are 10% and .2%. While speakers do produce -*s* at significantly higher rates for words ending in -*k*, the difference in rates is much smaller: an average of 8% -*s* plurals for -*k*-words, and 4% for all others. It appears that high variability in speaker behavior is the most relevant and consistent difference to how models handle plural generalization. We develop this point further in Chs. 6 and 7.

Marcus et al. observe that "frequency in the input to a pattern associator causes a greater tendency to generalize" 1995, 215; because of this, they argue, neural models should always prefer generalizing frequent plural classes instead the rare -*s*. If we consider the set of predictions for each individual stimulus item shown in Fig. 4.4, we can see that both the neural ED and the symbolic ATP do tend to generalize frequent classes. A majority of instances from each model assign the -*e* plural to the majority of stimulus items, while a mere handful of items receive a majority vote for the -*(e)n* or -*er* plural. This result, however, aligns well with speaker behavior: the majority of speakers assign the -*e* plural to 19 stimulus items, and -*(e)n* or -*er* to the remaining 5. It appears that speakers also tend to generalize frequent plural classes in the majority of cases.

We so see, however, a divergence between models and speakers — not necessarily in *which* plural classes receive majority assignments, but in *how strongly* the most frequent class is generalized. Fig. 4.7 shows the typical share of plural assignments to the most frequent class per stimulus item, from each model and from speakers. We can see that models typically agree on plural class assignments for a given stimulus item: across model instances, 80% or more on average will often assign the same plural class to a particular item. By contrast, speakers show much lower agreement — for any stimulus item, the most frequent plural class will be assigned by 45-50% of speakers on average. The divergence, then, is not that computational models generalize more frequent classes; it is that they generalize the most frequent class *more frequently*, yielding an item-level uniformity that we do not observe from speakers.

Figure 4.7 also highlights an interesting distinction between the models trained with and without gender. Both the ATP and the ED trained on wordform alone tend to agree more on plural class assignments for phonologically typical Rhyme stimuli, and show greater uncertainty for the phonologically atypical Non-Rhymes — including a higher proportion of rare *-s* plural assignments. For the ED trained with gender, however, we observe the inverse trend: more certainty for the phonologically atypical Non-Rhymes, and more diversity in predictions for the Rhymes. In particular, the other minority class *-er* appears most often on neuter nouns in the training corpus, so neuter gender presentation of the stimuli encourages the ED to generalize *-er* to stimuli with phonological neighbors in UniMorph. Although this interaction is not statistically significant under an ANOVA analysis of item-level entropy, the trend lends support to the interpretation presented in the preceding section: grammatical gender appears to substantially affect the phonological factors informing ED generalization, in this case shifting the category of stimuli for which the model is uncertain.

In any event, Fig. 4.7 illustrates that speaker plural generalizations are highly variable not only for phonologically atypical Non-Rhymes, but also for phonologically typical Rhymes. We might expect more plural class agreement on these stimuli: Rhymes have phonological neighbors in the existing German noun lexicon, and these neighbors would presumably influence plural class assignments in a more unified direction. Indeed, the ED and ATP trained on wordform alone show more agreement on phonologically typical Rhymes, likely for this reason. For computational models, phonological similarity to existing words facilitates confident plural class assignments. For speakers, we do not observe this: their plural generalizations are variable not only for phonologically atypical nouns, where uncertainty would be understandable, but also for nouns which are similar to familiar wordforms.

4.5 Conclusion

This chapter has reviewed behavioral and computational evidence relevant to Marcus et al.'s analysis of the German plural system. Their *minority default* theory posits that the plural suffix -*s* is a **rule-generated** (§2.1.1) default class, shown by its negatively-conditioned elsewhere distribution (§2.1.3). According to this analysis, German speakers generalize -*s* in heterogeneous default circumstances — such as to phonologically atypical nouns like *Bneik* — more often than its low **type frequency** would predict. In Study 1 (§4.2; McCurdy, Goldwater and Lopez, 2020), I reproduce this effect, finding that speakers reliably generalize -*s* more often to Marcus et al.'s phonologically atypical Non-Rhyme stimuli. In §4.3, however, I find that speakers use -*s* more often for the seven Non-Rhyme nouns like *Bneik* which end in -*k*; the other five phonologically atypical nouns like *Plaupf* and *Pröng* are not assigned -*s* more often than phonologically typical Rhyme nouns. Reanalysis of the original data (§4.3.2) indicates that this stimulus confound likely accounts for Marcus et al.'s initial finding. Although a neural ED trained with grammatical gender cannot reproduce this pattern (§4.2), neural and symbolic models trained on wordform alone successfully show speaker-like increased -*s* generalization to Non-Rhymes, also due to the -*k*-ending confound (§4.3.1). Considered as algorithmic linguistic theories (Baroni, 2022), these models provide a superior account of speaker -*s*-generalization behavior compared to the minority default hypothesis. Their predictions, however, are overconfident — i.e. show much lower within-item variability — compared to speakers (§4.4), and highly susceptible to grammatical gender. In Chs. 5 and 6, I assess whether grammatical gender shows a similarly strong influence on how speakers generalize the plural .

I conclude that we have no behavioral evidence that speakers systematically differentiate between phonologically typical and atypical nouns, or treat -*s* as a default class. Instead, they reliably default to *variation*, generalizing a balanced distribution of plural classes to both familiar and unfamiliar nouns. This distribution appears to track the lexical type frequencies of plural classes. In Chs. 6 and 7, I formally evaluate the extent to which speakers probability-match to lexical distributions.

Chapter 5

Speakers don't really use grammatical gender

5.1 Introduction

This chapter investigates how grammatical gender affects plural class generalization in German. In the previous chapter, I found that gender has a strong effect on how neural network models perform on this task: a recurrent neural Encoder-Decoder (ED) model trained *without* grammatical gender shows a speaker-like tendency to generalize the *-s* plural suffix to nouns ending in *-k*, while the same model architecture trained *with* grammatical gender does not capture this effect. In keeping with studies of other neural models (Goebel and Indefrey, 2000; Dankers et al., 2021; Beser, 2021), the ED appears highly sensitive to grammatical gender, preferring the conditional majority plural class within each gender — particularly the *-en* suffix for feminine nouns. This sensitivity is consistent with certain theoretical analyses of the German plural system. It clearly aligns with the rule-based gender default theory (§3.2.2); moreover, given the high mutual information between gender and plural class (Table 3.3), it also aligns with the predictability-based Low Conditional Entropy Conjecture (§2.1.2). The behavioral literature, however, provides conflicting evidence on whether speakers are similarly sensitive (§3.3.3).

Does grammatical gender show an equally strong influence on speakers' plural generalization? In this chapter, I answer this question with two behavioral studies using the same stimuli developed by Marcus et al.. Study 2 (§5.2; McCurdy, Lopez and Goldwater, 2020*a*) counterbalances grammatical gender presentation for all stimuli; Study 3 (§5.3; McCurdy, Lopez and Goldwater, 2020*b*) uses the same experimental design and offers participants additional financial incentives to converge on majority class productions. I find that, while speaker productions show weak gender effects, they are consistently far less sensitive to grammatical gender than predicted by the theoretical and computational models discussed above. Across both behavioral experiments, speaker productions are more closely correlated to the predictions of an ED trained on wordform alone, than to an ED trained with grammatical gender. Furthermore, the financial incentive manipulation in Study 3 has almost no impact on speakers' use of gender. I conclude that German speakers are at most mildly sensitive to grammatical gender in plural class generalization, and this outcome is incompatible with the strong gender effects predicted by certain theoretical and computational models.

5.2 Study 2

The material in this section has been published as McCurdy, Lopez and Goldwater (2020*a*).

5.2.1 Introduction

In recent years, neural models of natural language have proven to be powerful *statistical learners*, capable of representing linguistic patterns and the conditions under which they generalize to new forms (e.g. Kirov and Cotterell, 2018). Artificial language learning experiments show that humans are also statistical learners: when patterns appear consistently with certain cues in the input, speakers consistently rely on those cues to generalize patterns to new forms (Newport, 2016). Our research examines how two different statistical learners — neural encoder-decoder (ED) models and adult German speakers — use the cue of grammatical gender in plural inflection of novel words. Gender has a high statistical association with plural suffix: the feminine noun *Wahl* ("vote") is *Wahlen* in the plural, but the rhyming neuter noun *Mal* ("time") has the plural form *Male*. We expect that both speakers and the ED model will produce distributions over plural forms which are heavily conditioned on the gender of the input word. We find that the neural model is highly sensitive to grammatical gender; however, speaker productions appear more consistent with a distribution over plural suffixes which is *not* conditioned on gender. This result suggests that, though gender is highly informative to

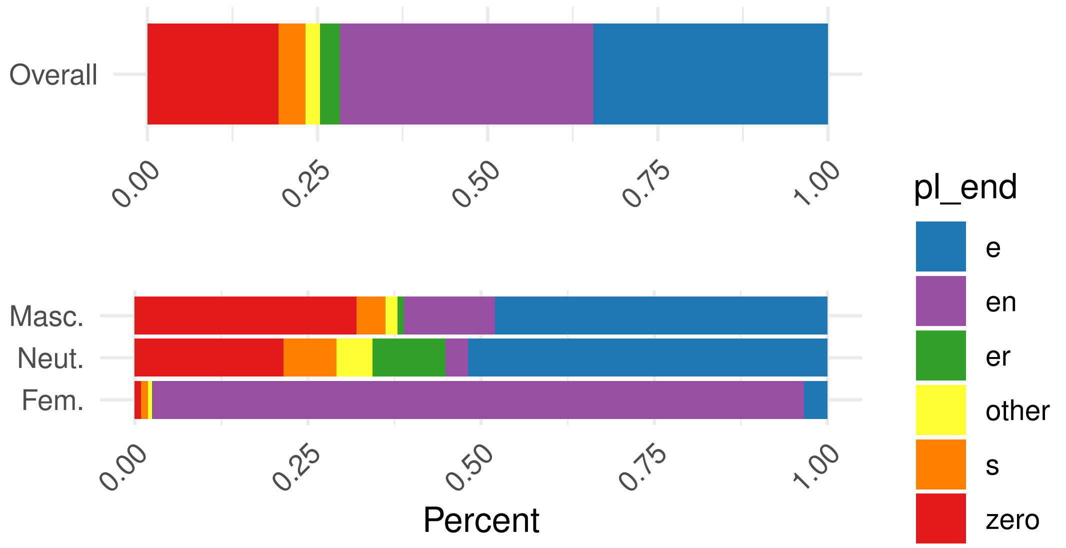


Figure 5.1: Distribution of plural suffix overall (upper) and by gender (lower) in the UniMorph corpus.

plural class, speakers may attend to different cues.

5.2.2 Background

Each German noun has two lexical attributes relevant to our analysis: its grammatical gender and plural inflection class. A noun can have masculine, neuter, or feminine gender, typically indicated by the preceding article. The other key lexical attribute, plural inflection class, is indicated by the plural form of the noun. See §3.1 for more extensive background.

Of the two most frequent plural suffixes, *-en* is highly associated with feminine nouns, and *-e* with nonfeminine (masculine and neuter). This statistical tendency is very strong (e.g. Table 3.2), leading some researchers to analyze these suffixes as gender-conditioned default classes (cf. §3.2.2, and §2.1.3 for discussion of linguistic defaults more broadly). While the psycholinguistic evidence in this area is mixed (§3.3.3), recurrent neural network models have shown robust sensitivity to grammatical gender in previous studies of plural generalization (§3.4).

We hypothesize that neural models and adult speakers are equally capable of using the information available from grammatical gender to predict number inflection. We expect both to demonstrate similar probability-matching conditional on gender (cf. §3.3.1) to the distribution shown in Figure 5.1 (lower), resulting in a majority use of *-en* for feminine nouns, and *-e* for masculine and neuter nouns.

5.2.3 Method

To compare how grammatical gender influences plural inflection for German speakers and neural models, we use a parallel production task on nonce words (a wug test) for both speakers and model. Our study largely follows the data collection and modeling procedures outlined in S1 (§4.2; McCurdy, Goldwater and Lopez, 2020).

Stimuli We use the 24 made-up nouns developed by Marcus et al. (1995) (Table 3.6). By design, these nouns lack strong phonological cues to plural class (c.f. §3.3.2). In their original study, Marcus et al. did not find a significant effect of grammatical gender; however, Zaretsky and Lange (2016) used the same stimuli and reported gender effects in the expected direction — participants used -*en* more on feminine nouns, and -*e* more for nonfeminine nouns. Zaretsky and Lange speculate that these discrepant findings stem from differences in the two study designs: scale (the earlier study had 48 participants, the later one 585) and task (acceptability ratings vs. elicited productions). A third differentiating factor is the presence of semantic cues in the Marcus et al. study, which provided sentence contexts around the nonce words; for example, a sentence like *Die grünen BRALS sind billiger* ("The green brals are cheaper") would imply that the nonce word *Bral* referred to an object, whereas *Die BRALS sind ein bißchen komisch* ("The Brals are a bit weird") would imply that *Bral* was a family name. As adult learners can attend to formal and semantic cues under different conditions (Culbertson et al., 2017), it's possible that this manipulation directed participant focus toward semantic cues rather than grammatical gender. Zaretsky and Lange provided no semantic context in their experiment, only presenting the indefinite article and word form to participants (e.g. *Ein Bral*, "a [masculine/neuter] bral"). Our experimental design for both speakers and the neural model is closer to that of Zaretsky and Lange (2016): we elicit plural form productions and provide no semantic cues. This suggests we might also expect to find a robust effect of grammatical gender for these stimuli.

Human data collection We collected production data from 92 native German speakers[1] through an online survey. Participants saw each noun in the singular with a definite article indicating grammatical gender (e.g. *Der Bral* for masculine, *Das Bral* neuter, *Die Bral* feminine), and typed a plural-inflected form. Participants were randomly assigned to one of three lists. Grammatical gender was counterbalanced within lists (each participant saw 8 feminine, 8 masculine, and 8 neuter nouns) and across lists (each noun appeared with a different gender in each list).

Encoder-decoder model We follow other recent work in using the architecture of Kann and Schütze (2016), which has been proposed for cognitive modeling (Kirov and Cotterell, 2018, cf. §2.3.3). For the task of German number inflection, the ED takes as input a character sequence representing the singular nominative form of a noun, preceded by a special character for grammatical gender (e.g. $\langle f \rangle$ W A H L; $\langle f \rangle$ indicates feminine, $\langle m \rangle$ masculine, and $\langle n \rangle$ neuter). The model is trained to produce the noun's corresponding nominative plural form as output (e.g. W A H L E N). We used the 11,243 German nouns in UniMorph (Kirov et al., 2016) as our corpus, and added noun gender by merging the dataset with another Wiktionary scrape.[2] Please see §4.2 for implementation details.

Following Corkery et al. (2019), we trained 25 separate random initializations of the same model architecture. This allows separate model instances to be treated as simulated "speakers", letting us aggregate productions and compare more directly to human speaker data. For evaluation, we combined each of the 24 noun stimuli with each of the three grammatical genders, and provided the resulting 72 items as input to each model instance.

5.2.4 Results

Our results (Figure 5.2) show that both speakers and the ED model are sensitive to grammatical gender, but the model relies on this cue considerably more than speakers. Statistical analysis confirms that a) both speakers and the model show reliable effects of grammatical gender on their plural form productions, and b) gender effects are substantially greater for model productions. We fit two separate mixed-effect binomial logistic models using the lme4

---

[1] Participants were recruited through the platform Prolific. Of 100 tested, 8 were excluded for failing attention checks.

[2] `https://github.com/gambolputty/german-nouns/`

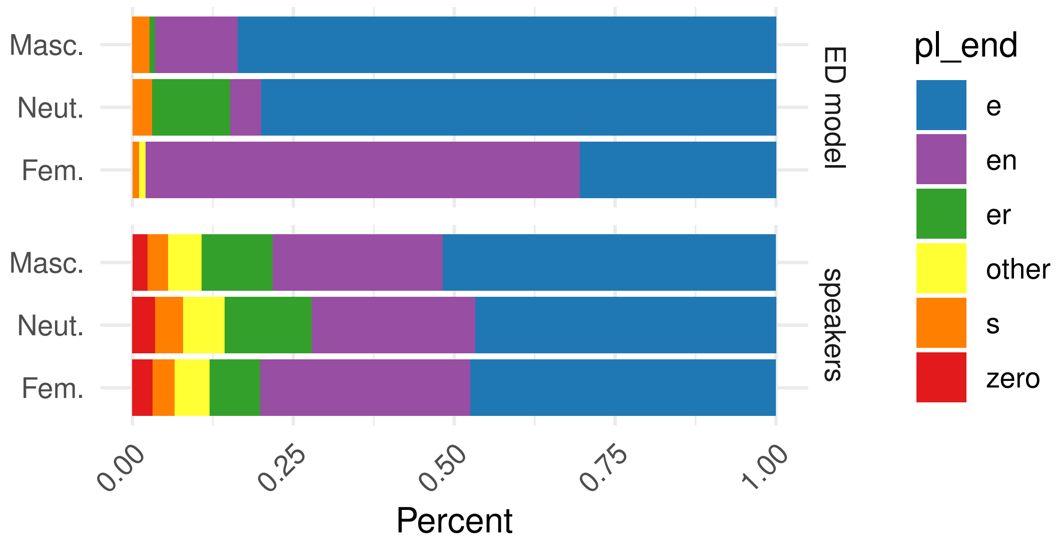


Figure 5.2: Plural suffix productions by gender, speakers (lower) vs. ED model (upper)

package (Bates et al., 2015) in R (R Core Team, 2019). Item (i.e. stimulus word) and subject (participant for human study, random seed for ED model) were included as random effects. Both models were fit using a stepwise procedure. We started with a baseline model of intercept plus random effects and incrementally added the following fixed effects (with sum-coded contrasts): grammatical gender (masculine coded as 1, neuter as 2, feminine not contrasted), data source (ED model coded as 1, speakers not contrasted), gender by source interaction. Each additional fixed effect produced a significantly improved fit as measured by a chi-squared test.

The final model for both *-e* and *-en* production includes all fixed and random effects described above. For both plural suffixes, model results indicate a significant main effect of gender from both speakers and the ED model, and a significant interaction with data source, corresponding to a stronger effect of gender from ED model productions. For *-e* productions, there is also a main effect of data source: the ED model reliably produces *-e* more than speakers do overall. The *-en* model shows no significant main effect for source. When model predictions are transformed to responses and fit to the original data, the binomial model of *-e* production achieves an overall predictive accuracy of 75% (precision 0.77, recall 0.79, F1 0.78), while the *-en* model has 82% predictive accuracy (precision 0.71, recall 0.59, F1 0.65).

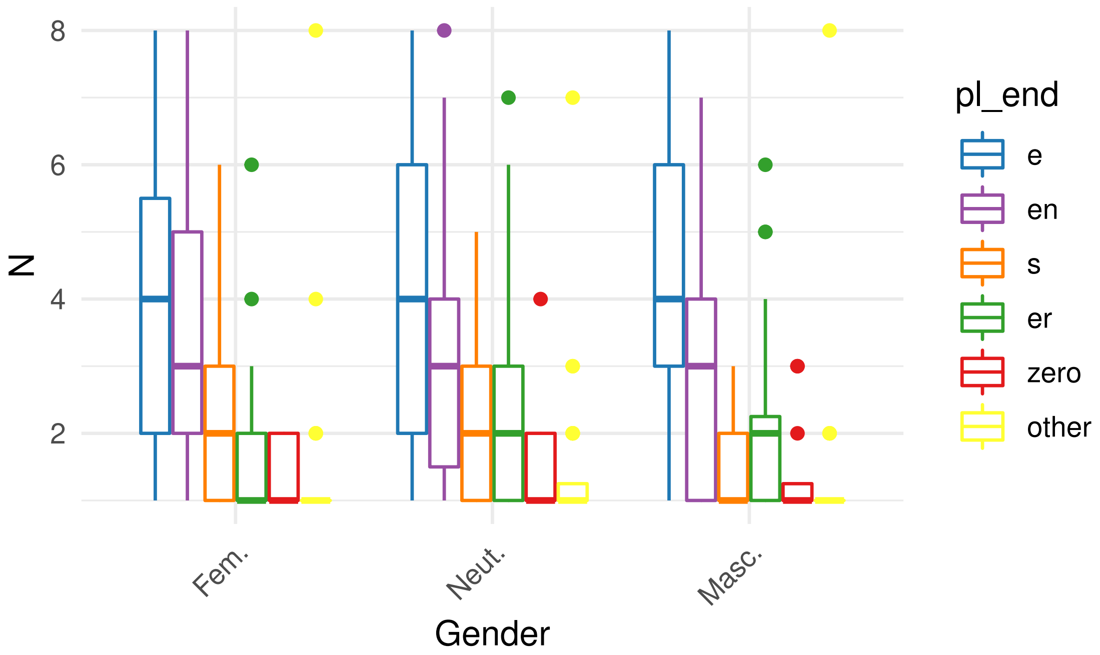


Figure 5.3: Individual speaker variation in plural suffix production by gender. Each speaker saw 8 words from each gender, shown on the y-axis. For each gender and plural suffix, the boxes indicate the median and interquartile range of individual speaker productions for that combination. For all gender categories, the median number of -*e* productions is 4, while the median number of -*en* productions is 3.

Sanity checks As human speakers show high inter-participant variability on this task (Fig. 5.3), we performed additional separate analysis on the speaker data.[3] We fit the same model as previously described, with the exception that the data source factor was omitted, as all data came from speakers. We also fit models using Masculine and Neuter as the reference gender in the sum contrast coding scheme, to see whether they yielded different results from the original model's Feminine reference level.

The speaker-only model shows a reduced but consistent effect of gender (Tab. 5.1). Speakers reliably produce -*en* more for feminine nouns, and less for neuter nouns, relative to the grand mean. Speakers also reliably produce -*e* more for masculine nouns. These difference are statistically significant even though, for all three genders, speakers produce -*e* more than -*en* (Fig. 5.3).

Intriguingly, the speaker productions are not only less sensitive to grammatical gender, they also appear very consistent with the overall type frequency distribution of the plural suffixes, *unconditioned* on gender. To quantify this intuition, we looked at how the distribution of plural

---

[3] We thank an anonymous reviewer for highlighting this issue and suggesting these validity checks.

Suffix	Effect	Fem.	Neut.
-en	gdr.fem	.	.31 (.08) ***
	gdr.neut	-.19 (.08) *	.
	gdr.masc	-.12 (.08)	-.12 (.08)
-e	gdr.fem	.	-.07 (.07)
	gdr.neut	-.1 (.07)	.
	gdr.masc	.17 (.07) *	.17 (.07) *

Table 5.1: Speaker data only: statistical effect of different gender reference levels in contrast coding. Header shows reference level, cells show estimated coefficient (standard error in parentheses). Estimates for Masculine reference level are identical to rows already shown (e.g. suffix *-en*: $-.19$ for gdr.neut, $.31$ for gdr.fem). Stars indicate significance level: * ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001.

	Overall-TF	Gender-TF	ED
Speakers	**.67**	.49	.49
	(.60, .72)	(.40, .56)	(.35, .61)
ED	.41	**.62**	
	(.27, .54)	(.50, .71)	

Table 5.2: Correlations (Pearson's *r*, 95% confidence intervals in parentheses below) between item-level production percentages for speakers and ED model with 1) overall type frequency (Overall-TF), 2) gender-conditioned type frequency (Gender-TF), 3) each other.

suffixes produced over each of the 72 noun-gender item combinations *correlated* to various other metrics. We asked three questions: 1) How well do item-level speaker and ED model productions correlate with *each other*? 2) How well do both sets of item-level productions correlate with the *gender-conditioned distribution* of plural suffix types observed in the German lexicon? 3) How well do both sets correlate with the *unconditioned overall distribution* of types? Table 5.2 shows the results: while item-level ED outputs are most correlated with the gender-conditioned distribution, item-level speaker data is most correlated[4] with the overall (unconditioned) type frequency.[5] Even though the speaker and ED data are matched by item, their productions have a lower correlation with each other than with the general type-frequency distributions.

---

[4] Table 5.2 shows results from Pearson's linear correlation; analysis with Spearman's rank correlation coefficient showed the same trend.

[5] To avoid potential confounds from both training and measuring on the UniMorph corpus, our estimates of gender-conditioned and overall type frequency are derived from Zaretsky et al. (2013)'s analysis of the thousand most frequent nouns from the DeReWo corpus.

	Overall-TF	Gender-TF
Speakers	**.78** (.73, .82)	.70 (.64, .75)
ED	.47 (.33, .59)	**.71** (.61, .78)

Table 5.3: Correlations (Pearson's r, 95% CI in parens) between item-level production percentages for speakers and ED model with 1) overall type frequency (Overall-TF), 2) gender-conditioned type frequency (Gender-TF), *only* considering consonant-final monosyllabic nouns in UniMorph (shown in Figure 5.4).

5.2.5 Discussion

We hypothesized that adult speakers and neural encoder-decoder models would make similar use of grammatical gender when inflecting novel words in the plural, as gender is a salient and consistent cue to plural inflection class, especially in an experimental setup where semantic cues are absent. Contrary to expectations, our results indicate that both learners attend to grammatical gender, but to different degrees — the neural model is much *more* sensitive to grammatical gender than adult speakers, whose productions are closer to the overall type frequency of plural suffixes in German.

The neural model's use of grammatical gender is not surprising, as it aligns with earlier findings (c.f. Goebel and Indefrey, 2000); however, the speakers' *lack* of attention to gender is unexpected. In their large-scale production study with the same noun stimuli, Zaretsky and Lange (2016) found reliable effects of grammatical gender: their participants used *-en* for 33% of feminine nouns, versus 19% of non-feminine nouns (compare to our study: 33% vs. 26%). *-e* also appeared more with nonfeminine nouns (49% vs. 41%), although the effect was not statistically significant. Nonetheless, they note that *-e* was most frequently produced for feminine nouns as well as nonfeminine nouns, consistent with our results, and their data shows a similarly broad distribution over types. Despite other differences between our study design and theirs (e.g. online vs. in-person data collection, typed vs. written modality, German speakers from various backgrounds vs. one region), we consider our results fundamentally aligned: speakers show a slight but statistically reliable effect of gender on *-en* and *-e* production, in both cases much less than the effect shown by the ED model.

One possibility is that the phonological forms of our noun stimuli provide their own statistical conditioning, to a stronger degree than anticipated. This is illustrated in Figure 5.4, which plots the distribution of nouns in UniMorph sharing two key properties with our stimuli: they are monosyllabic and end in a consonant. On the one hand, nouns with this type of

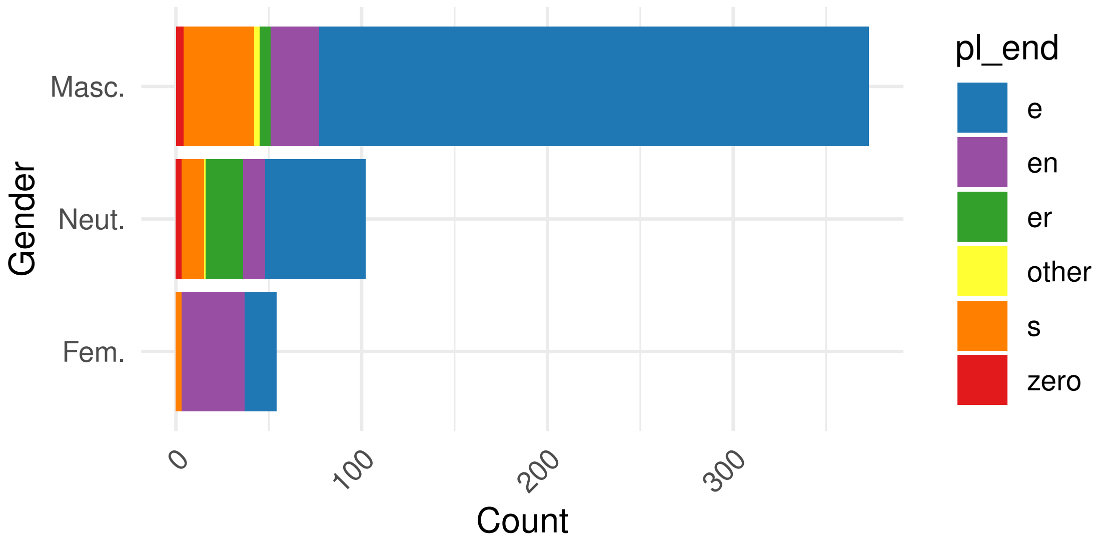


Figure 5.4: Distribution of plural suffix by gender for consonant-final monosyllabic words in UniMorph

form clearly also show gender conditioning, with *-en* much more prevalent among feminine nouns. On the other hand, nouns with this general form are predominantly masculine gender, and the numerical prevalence of nonfeminine forms may diminish speakers' sensitivity to a rare feminine gender cue, such as they encounter in our experiment. Under this account, adult speakers condition their plural productions upon phonological form to a greater extent than grammatical gender. The results in Table 5.3 further support this interpretation. Looking only at the consonant-final monosyllabic words plotted in Figure 5.4, ED model productions show a higher correlation to the gender-conditioned distribution over plural suffixes, while the highest correlation generally (.78) appears between speakers productions and the overall distribution of plural classes for these phonologically similar words. The potential shortcoming of the ED as a cognitive model, then, is that it assigns too much weight to the cue of grammatical gender, even though it is statistically reasonable to do so.

In conclusion, our comparison of neural encoder-decoder models and adult German speakers found a significant difference in their use of grammatical gender as a cue to plural inflection. Although this cue is highly informative, speakers — unlike neural models — appear relatively insensitive to gender in our task. This finding suggests that speakers may attend more readily to other cues such as phonology, and therefore match productions to a different distribution which shows less gender conditioning.

In dieser Studie werden Sie statt echten Wörtern (wie z.B. "Messer") fiktive Wörter sehen, die andere Teilnehmer auch schon gesehen haben. Für jedes Wort, **vermuten Sie bitte, welche Pluralform am häufigsten von anderen Teilnehmern angegeben wurde**. Bei jeder richtiger Vermutung verdienen Sie noch 2 Cent zusätzlich als Bonus.

In this study, rather than real words such as "knife," you will see made-up words which have already been seen by other participants. For each word, **please guess which plural form was given most frequently by other participants**. You will receive 2 cents additionally as a bonus for each correct guess.

Table 5.4: German-language instructions for Study 3 and their English translation.

5.3 Study 3

The main finding of Study 2 is the unexpectedly weak effect of grammatical gender on speaker productions, which stands in contrast to the strong effect of gender on the neural Encoder-Decoder (ED) model predictions. One possible explanation is a design flaw in the behavioral experiment: perhaps speakers simply had no compelling reason to attend to grammatical gender while completing the study. This possibility motivated a follow-up experiment to address this limitation.

In an interactive artificial language learning study, Perfors (2016) found that financial incentives could systematically affect adult speakers' regularization behavior. Participants operated in pairs, and learned to map words to specific objects. In one condition, both participants in a given pair were compensated based on the *accuracy* of the mappings they learned, evaluated by an objective standard; participants in this condition tended to probability-match. Other participants, however, were compensated instead based on how closely they *matched* each others' learned mappings. In this condition, participants showed significantly more regularization. Perfors notes that this result aligns with earlier findings of financial incentives leading participants to reduce probability-matching behavior (Vulkan, 2000; Shanks et al., 2002). Based on this finding, I extend the design of Study 2 to include financial incentives for participants to match others' responses. These results have been presented at the 26th Architectures and Mechanisms for Language Processing (AMLaP) conference (McCurdy, Lopez and Goldwater, 2020*b*).

Study	$H(C)$	$H(C \mid G)$	$MI(C;G)$
S2	1.928	1.920	0.008
S3	1.930	1.918	0.013

Table 5.5: Mutual information (MI) between gender and plural class for Study 2 and Study 3. The additional manipulation in S3 produces only a small MI increase.

5.3.1 Behavioral experiment

Study 3 uses the same design and materials as Study 2, with one additional manipulation: participants were offered financial incentives for producing the same plural form as the majority of other participants. Table 5.4 shows the updated instructions. While two cents may not appear a significant financial incentive, a participant who produced the majority plural form each time would earn an additional 48 cents — which, given the short amount of time spent on the study (median time to completion was roughly 7 minutes), would represent a 35% increase in their earnings for the study.

Note that this design is not directly parallel to the experiments considered above, where participants typically receive direct feedback on their responses; for instance, the participants in Perfors' match condition could observe how well their productions aligned with those of their partner. In contrast, speakers receive no feedback about the behavior of others in the course of completing this experiment; the only reward signal comes in the form of the bonus, which is calculated and distributed after the experiment. Nonetheless, the reasoning behind this design choice is that speakers may be motivated to use additional information available to them to increase their reward. Given the strong statistical relationship between grammatical gender and plural class, they might reason that *other* speakers condition their responses on the information available to them — i.e. grammatical gender — and therefore attend more reliably to this cue.

100 German speaking participants were recruited on Prolific for this study. Participants who had taken part in either Study 1 or Study 2 were excluded.

5.3.2 Results

The addition of financial incentives in S3 led to only a very slight increase in use of grammatical gender. This is visible in Table 5.5, which shows a small increase in mutual information

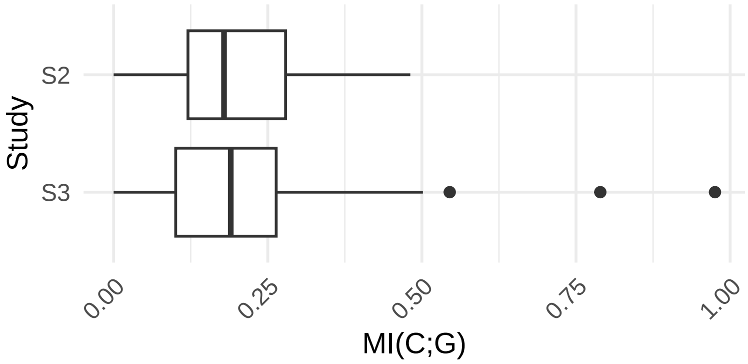


Figure 5.5: Distribution of mutual information (MI) between plural class and grammatical gender per participant per study. Note that the slight increase in MI for S3 appears largely driven by three individual participants.

	ED-gender	ED-no-gender
Speakers	.56 (.42, .67)	**.73** (.65, .80)

Table 5.6: Correlations (Pearson's *r*, 95% CI in parens) between item-level production percentages for speakers and ED model *with* and *without* explicit grammatical gender indicated. Correlation was evaluated across three bins per item and gender: percent -*e* produced, percent -*en* produced, and percent all other productions. Item-level speaker data is more correlated to the productions of the model without gender than the model with gender.

between plural class and grammatical gender for S3 relative to S2. Figure 5.5 shows the distribution of this quantity over individual participants in each experiment, revealing that only 2 or 3 individual participants in S3 show any evidence of using gender more reliably than the participants in S2.

Study 2 compared participant productions to two lexical baselines — overall type frequency, and gender-conditioned type frequency — and one model: the Recurrent Neural Network Encoder-Decoder (ED) trained *with* grammatical gender. In that study, I found that speaker productions showed a higher correlation to the lexical type frequency baseline which was *not* conditioned on gender, than to either the ED predictions or the gender-conditioned lexical baseline.

Given the high level of consistency between speaker behavior in S2 and S3, I combine the two experiments to compare how closely they match the productions of the ED model trained *without* grammatical gender, discussed in §4.3.1. The results of the correlation analysis are presented in Table 5.6. Speaker productions overall show a higher correlation to the predictions of the ED trained on wordform alone.

5.4 Discussion

The key result of the two behavioral studies reviewed here is speakers' robust lack of sensitivity to the grammatical gender of these novel noun stimuli in plural generalization. To be clear, gender *does* show a statistically significant impact on speakers' plural productions, as reported in Table 3 of §5.2.[6] The effect, however, is much weaker than expected based on either neural ED model predictions (c.f. Table 3 in §5.2) or the German noun lexicon as a whole (compare the reported mutual information in Table 5.5 to Table 3.2).

The largely ineffective experimental manipulation in S3 illustrates that this lack of sensitivity to grammatical gender is a relatively robust effect. Of course, it could also indicate further issues in experimental design. Speakers may not have responded to the financial incentives because the amounts were too small, or perhaps the lack of feedback during the experiment prevented speakers from evaluating and updating their behavior. In Chapter 6, Study 4 addresses these issues with an alternative task design: speakers must assign both grammatical gender and plural inflected form to the novel noun stimuli, so the task necessitates at least a minimal level of attention to grammatical gender.

This lack of sensitivity to grammatical gender is compatible with some previous behavioral evidence (§3.3.3), but incongruous with some theoretical analyses (§3.2). These results are clearly not readily compatible with the gender-based default analysis of the German plural system (§3.2.2). Advocates of this theory, however, have typically applied it to the task of linguistic description rather than behavioral prediction, so this may not be an appropriate evaluation; the statistical relationship between gender and plural class does not appear to drive speakers' initial generalization behavior, but it may emerge in the lexicon over time, possibly subjected to interactive pressures during cultural transmission as described in §2.2.2. These findings are perhaps more directly incompatible with the broader predictability account of morphological generalization (§2.1.2), which takes a more functional view explicitly connected to speaker behavior, and has informed some behavioral studies (§2.2.2). The Low Conditional Entropy Conjecture (Ackerman and Malouf, 2013) predicts that speakers will draw upon all sources of information which can reduce uncertainty in morphological generalization, including grammatical gender. The results of S2 and S3 do not accord with this interpretation.

---

[6] The same significance relation holds when S3 data is included as well, reported in McCurdy, Lopez and Goldwater (2020*b*).

In contrast to speakers, the neural Encoder-Decoder model treats grammatical gender as a predictable cue to the plural class of unknown German nouns. This computational result is consistent with other modeling studies. Goebel and Indefrey (2000) find that their simple Recurrent Neural Network (RNN) model is highly sensitive to grammatical gender, leading them to propose a gender-based default analysis of the German plural system (see also Indefrey, 1999). Dankers et al. (2021) conduct a detailed causal and representational analysis of an RNN model with an additional memory cell (a Long Short-Term Memory network, or LSTM; Hochreiter and Schmidhuber, 1997) trained on the same task. They find a similar reliance on grammatical gender: analysis of the memory cells and hidden states shows that the model learns to predict the *-en* feminine class immediately upon encountering the feminine article *die*. Although it is far from a proper evaluation, this strong sensitivity to grammatical gender supports the conjectural explanation for the asymmetric results discussed in in Chapter 4. Recall that the neural model trained *without* gender showed speaker-like increases in *-s* production for nouns ending in *-k*, but the model trained *with* gender failed to show this effect. The findings reviewed here suggest that neural models find grammatical gender such a strong predictive signal that it may overpower more subtle associations, like the statistical connection between a rare noun ending *-k* and a rare plural class *-s*. This asymmetry is noteworthy in the context of modern deep learning research overall, where human-like linguistic behavior is generally achieved through more data, not less; there do not seem to be many other cases in the literature where censoring data, or parts of the data signal, improves performance. It is also noteworthy that this behavioral failure arises for a cue which is so robust, that both theoretical linguists and statistical models agree on its significance. To consider the ED as an algorithmic linguistic theory following Baroni (2022), it appears to be more successful at predicting behavior when it diverges from formal linguistic theory — as for the *-s*-*k* connection discussed in Chapter 4 — than when it converges, as in the case of grammatical gender.

So why do speakers disregard grammatical gender, a cue that neural network models rely on to generalize plural class? One possibility comes from the Functional Theory of Gender Paradigms proposed by Dye et al. (2017). They argue that grammatical gender in fact plays a crucial role in predictability, but that role can only be understood in the wider context of the *sentence*. They support this interpretation with an information-theoretic analysis of German nouns in a large corpus, and find that gender marking has the aggregate effect of modulating nominal entropy at the sentence level: low-frequency nouns which would be difficult to predict in a particular sentence context become more predictable when preceded by a gendered article, facilitating increased lexical diversity through the use of less frequent nouns. This finding suggests that grammatical gender *does* crucially contribute to predictability, but that contribution occurs in the sentence-level context of ordinary language use, rather than in the mapping from one morphological paradigm cell (i.e. singular) to another (plural). Ramscar (2021) further develops this argument, stating that psycholinguistic and language acquisition research should properly focus on how inflectional morphology is learned and represented in this more realistic context, instead of the relatively artificial scenario of mapping within morphological paradigms as I have pursued here. In my opinion, this critique points toward useful future research directions and modeling work focused on broader linguistic contexts; however, given that speakers show fairly robust and consistent behavior (i.e. variable, aligned with overall type frequencies, insensitive to grammatical gender) on the relatively artificial task of mapping novel nouns to plural forms, I believe that characterizing and modeling this generalization behavior remains a valid line of scientific inquiry.

5.5 Conclusion

This chapter has presented findings from two studies, S2 and S3, which use the same stimuli as in S1 (Ch. 4) except with counterbalanced grammatical gender presentation. S3 includes additional financial incentives aimed to motivate participants to incorporate all available information when generalizing plural forms, with the reasoning that this would encourage them to rely more upon grammatical gender as a highly predictive cue; however, this intervention yielded only a slight increase in the target behavior. I conclude that speakers are consistently insensitive to grammatical gender in plural class generalization. This raises issues for the predictability view of morphological regularity, especially when speaker behavior is compared to neural network models which reliably use grammatical gender to predict plural class: computational models may use all available information to reduce uncertainty in generalization, but speakers apparently do not.

Chapter 6

Speakers probability-match and condition gender on phonology

6.1 Introduction

In the previous chapter, I found that German speakers are only mildly sensitive to grammatical gender in plural generalization (Study 2, §5.2), even when offered financial incentives to generalize more predictably (Study 3, §5.3). This chapter reports Study 4 (McCurdy et al., 2022), a follow-up experiment in which speakers are asked to produce both the grammatical gender and the inflected plural form of the same novel stimuli. This task design requires participants to explicitly attend to grammatical gender, and its joint distribution with plural forms. If speakers are at all aware of the statistical relationship between these two variables, the task design in S4 should support this with behavioral evidence.

The results of S4 further support the lack of sensitivity to grammatical gender observed in Chapter 5. I find that speaker productions do not reflect the relatively high amount of mutual information between gender and plural class in the lexicon as a whole, but rather the lower level of mutual information found in a specific subset of the lexicon — namely, nouns which share the monosyllabic structure of the experimental stimuli. I conclude that speakers probability-match to a lexical distribution conditioned on phonology, specifically syllabicity.

6.2 Study 4

The contents of this section have been published as McCurdy et al. (2022).

6.2.1 Motivation

Research in artificial language learning, reviewed in §2.2.2, shows that adult speakers have a range of responses to unpredictable inconsistencies in their linguistic input. Under some circumstances, they *probability-match* and reproduce the variation in their input distribution, while in other circumstances they prefer to *regularize*[1] and produce more consistent patterns (Hudson Kam, 2019). Often, speakers regularize by increasing production of the most frequent variant in their input (Hudson Kam and Newport, 2005, 2009). The learning biases influencing speaker behavior in these experiments are not fully understood, and show complex interactions with communicative pressures in cultural transmission (Smith et al., 2017). As a result, it is challenging to anticipate which artificial language findings will apply in more complex natural language environments, such as the German plural inflection task we explore in this study.

Research on natural language variation shows that it is typically *conditioned* upon multiple factors, both linguistic (e.g. phonological environment) and non-linguistic (e.g. speaker identity) (Chambers and Schilling, 2018). Conditional variation provides another mechanism for regularization: unpredictable variation can become predictable when conditioned on particular linguistic contexts. Table 2.6 gives illustrative examples of regularization, probability-matching, and conditional regularization.

Note that our example artificial language experiment frames variation with respect to static attributes within a lexicon: each individual noun has two fixed classes (expressed by the article and the plural form), and we consider how speakers might use membership in one class (e.g. article) as a cue to signal membership in another class (plural form). Artificial language learning studies have shown that adult speakers can learn to condition noun class assignment on such markers when they are statistically reliable (Frigo and McDonald, 1998). Culbertson et al. (2017) found that learners may prefer different cues to noun class (e.g. phonological vs. semantic cues) based on salience or early availability in training. While they found reliable statistical main effects from their experimental cue manipulation, their data show a broad range

---

[1] N.B. we use the term *regularize* in the linguistics sense (reduce variation), not the machine learning sense (reduce overfitting).

of variation within individuals as well, suggesting the type of variation in speaker strategies illustrated by the hypothetical case in Table 2.6.

The studies discussed above explore speaker generalization using toy lexicons, where the amount and type of variation can be manipulated experimentally. However, in principle it should be possible to apply some of the same analysis methods to the more complex case of generalization from natural language. In particular, German number inflection provides a complex natural-language test case for the type of lexical variation seen in our hypothetical experiment. Some aspects of the German plural system are well-described by rules (e.g. derived nouns; Augst, 1979), but other parts of the lexicon show more complex probabilistic relations, and psycholinguistic experiments reveal considerable variation between speakers when they are asked to produce the plural forms of novel words (Mugdan, 1977; Köpcke, 1988; Zaretsky and Lange, 2016; McCurdy, Goldwater and Lopez, 2020).

In this work, we adopt the framework of probability-matching versus regularization to shed light on this variability. We ask whether variation in German number inflection of novel words can be explained in terms of a) lexical statistics and b) variation in individual speaker strategies. Do speakers predominantly probability-match to the distribution observed in the lexicon, leading to the variation observed in behavioral experiments? Or do they predominantly regularize, but with different speakers pursuing different strategies (e.g. reducing conditional vs. overall variation) which lead to a general appearance of inconsistent behavior?

We use the information-theoretic definition of regularization presented by Ferdinand et al. (2019) to evaluate individual behavior in terms of entropy. We take the joint distribution of grammatical gender (G) and plural inflection class (C) observed in the lexicon as a reference distribution to assess German speaker behavior on a dual task: for each of 24 novel nouns, identify its grammatical gender, and produce its plural inflected form. We find that, consistent with some artificial language experiments, adult speakers largely probability-match the conditional variation observed in the input, and disregard an alternative strategy of gender-conditioned regularization. Our work shows that lexical statistics across items can predict speaker behavior within novel items, connecting artificial language findings with natural language behavior.

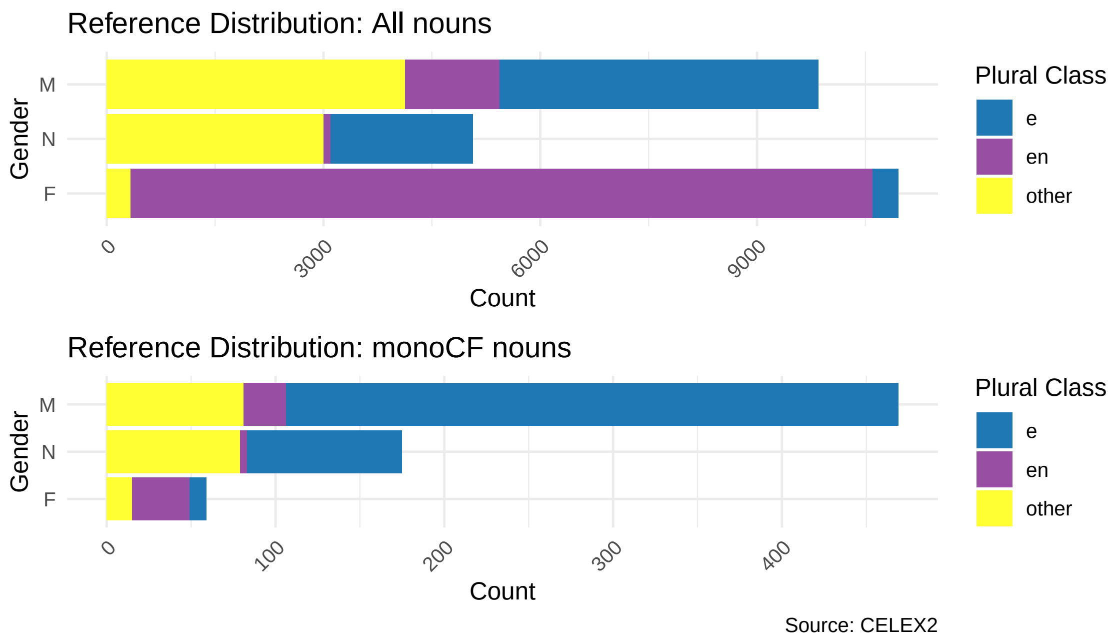


Figure 6.1: Reference distributions calculated from German CELEX2 (Baayen et al., 1995). Top, all nouns (excluding ambiguous gender nouns). Bottom, monosyllabic consonant-final nouns.

6.2.2 Background

German number inflection Each German noun has two lexical attributes relevant to our analysis: its grammatical gender (G) and plural inflection class (C). A noun can have masculine (M), neuter (N), or feminine (F) gender, and this lexical property has a complex relation to the noun's phonology and semantics (e.g. Köpcke and Zubin, 1984). The other key lexical attribute, plural inflection class, is indicated by the plural form of the noun. For more background on the German plural system, see §3.1.

As regularization often involves increasing frequent variants, we focus on the two most frequent plural suffixes and their relationship to grammatical gender. Fig. 6.1a shows the joint distribution of 3 simplified plural class (-*e*, -*en*, and "other") by gender over **all nouns** in CELEX2. Fig. 6.1b focuses on the subset of nouns in CELEX2 with a similar phonological shape to our experimental stimuli, i.e. monosyllabic and consonant-final (**monoCF**).[2]

---

[2] N.B. In this dissertation section, I refer to the reference distribution as 'monoCF' to align with the analysis and graphics presented in (McCurdy et al., 2022). In practice, however, as the vast majority of German monosyllabic nouns are also consonant-final, I use the more clear term 'monosyllabic' for this reference baseline elsewhere in the dissertation.

	$H(G)$	$H(C)$	$MI(C;G)$	$U(C \mid G)$
All nouns	1.52	1.54	0.61	40%
All (6 cl.)	1.52	1.98	0.67	34%
monoCF	1.19	1.21	0.18	14%
mCF (6 cl.)	1.19	1.55	0.23	15%

Table 6.1: CELEX2 entropy measurements for gender $H(G)$, plural class $H(C)$, mutual information between plural class and gender $MI(C;G)$, and percentage plural variation explained by gender $U(C|G)$. We see similar values whether using our simplified 3-class analysis or a more traditional 6-class analysis for C.

Gender and plural class Our key research question is whether German speakers will regularize overall variation, probability match the observed lexical distribution, or regularize conditional variation. In the latter case, grammatical gender is the most viable option on which to condition plural class, for several reasons. 1) There is a clear strong statistical relationship between gender and plural class, evident in Fig. 6.1. Williams et al. (2020) analyze a subset of German nouns in CELEX2, and estimate that 25% of the variation in inflection class (including all plurals and cases) can be explained by grammatical gender. For our simplified set of inflection classes, we estimate 40% (Table 6.1; see also Tables 3.2 and 3.3, and discussion in §3.1). 2) Many linguists have analyzed grammatical gender as the primary determinant of plural class, with *-e* as the default class for non-feminine nouns, and *-en* for feminine nouns (e.g. August, 1979; Wiese, 1999; Bittner, 1999, and further discussion in §3.2.2). 3) Neural models of German inflection reliably learn to condition plural class on gender (Goebel and Indefrey, 2000; McCurdy, Lopez and Goldwater, 2020*a*; Dankers et al., 2021, and discussion in §3.4). Despite this, many psycholinguistic studies report little (e.g. Köpcke, 1988; Zaretsky and Lange, 2016; McCurdy, Lopez and Goldwater, 2020*a*) or no (e.g. Mugdan, 1977; Spreng, 2004) effect of gender on speaker productions; see discussion in §3.3.

Regularization and German plurals Investigating regularization leads us to focus on the two most frequent inflection classes. This contrasts with some of the literature on German plural inflection; for instance, the influential Dual Mechanism Model (e.g. Marcus et al., 1995; Clahsen, 1999*b*, and discussion in §3.2.1) emphasizes the regularity the minority class *-s*. Other linguistic analyses of the German plural system have focused on productivity and type frequency (e.g. Köpcke, 1988; Bybee, 1995; Yang, 2016; Heitmeier et al., 2021). For an extensive review of the relevant linguistic literature, please refer to §3.2.

Herce (2019, c.f. §2.1.1) notes that the term "regularity" is associated with many distinct concepts in the linguistics literature, and recommends that researchers use more precise language, e.g. "productivity" or "predictability." Our approach emphasizes the "predictability" dimension, in line with other recent attempts to formalize an information-theoretic concept of morphological regularity (e.g. Ackerman and Malouf, 2013; Cotterell, Kirov, Hulden and Eisner, 2018; Wu et al., 2019, and discussion in §2.1.2). Note, however, that these analyses use the lexicon to estimate the *regularity* of a lexical item, for example to predict that the English past tense form "jumped" is more regular (i.e. predictable) than "ran." In contrast, we use the lexicon to assess *regularization* behavior by speakers: do they maintain the level of variation present in the lexicon, or introduce more predictability to novel lexical items? See §2.2 for further review and analysis of the relationship between regularity and regularization.

6.2.3 Methods

Quantifying regularization

Ferdinand et al. (2019) present a novel quantitative analysis of regularization in terms of entropy. Under their definition, speaker regularizing behavior is formalized as the degree of entropy reduction relative to a reference distribution. All measures here originate with Shannon (1948). See §2.1.2 for a more general discussion of information-theoretic approaches to morphological analysis.

The first key measure is Shannon entropy, which quantifies in bits the complexity, or variation, over the distribution of a single categorical variant (c.f. Eq. 2.1). In our case, we're interested in entropy over plural class C:

$$H(C) = -\sum_{c \in C} \mathrm{P}(c) \log_2 \mathrm{P}(c) \tag{6.1}$$

Similarly, $H(G)$ gives the entropy of the distribution over grammatical gender.

The second key measure is conditional entropy (c.f. Eq. 2.4), which calculates the entropy of our variant of interest C conditioned on grammatical gender G:

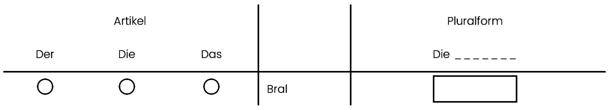


Figure 6.2: Task presentation for one item. To the left of the novel noun is gender selection, to the right, written plural.

$$H(C \mid G) = -\sum_{g \in G} \mathrm{P}(g) \sum_{c \in C} \mathrm{P}(c \mid g) \log_2 \mathrm{P}(c \mid g) \tag{6.2}$$

Subtracting conditional entropy from Shannon entropy gives the mutual information between the two variables:

$$MI(C;G) = MI(G;C) = H(C) - H(C \mid G) \tag{6.3}$$

The mutual information can be normalized by the Shannon entropy to estimate of the percentage of variation explained by the conditioning variable, known as the uncertainty coefficient (Williams et al., 2020)[3]:

$$U(C \mid G) = \frac{MI(C;G)}{H(C)} = \frac{H(C) - H(C \mid G)}{H(C)} \tag{6.4}$$

Under Ferdinand et al.'s framework, any reduction in entropy relative to the reference distribution qualifies as regularization. They note that this can be accomplished in three ways: reducing variation in either the distribution of the variant $H(C)$, or of the context $H(G)$, or the conditional distribution $H(C|G)$ (equivalent to increasing $MI(C;G)$).

Behavioral experiment

Stimuli The stimuli used in this experiment can be seen in Table 3.6. They comprise 24 monosyllabic nouns ending in a consonant (i.e. monoCF nouns), originally developed by Marcus et al. (1995). As seen in Fig. 6.1 and Tab. 6.1, this class of nouns is ambiguous in terms of plural class and grammatical gender. This makes them good candidates to assess regularizing

---

[3] We thank an anonymous reviewer for noting that mutual information is typically normalized with respect to the smaller entropy, in this case $H(G)$. We use $H(C)$ in the denominator as we are specifically interested in $U(C \mid G)$, the fraction of plural class entropy explained by gender, rather than the inverse relation $U(G \mid C)$.

behavior — other phonological classes of German nouns already have fully predictable inflection class assignments, e.g. nouns ending in schwa near-universally take the *-en* plural. These stimuli have also been used in multiple previous experiments (e.g. Marcus et al., 1995; Zaretsky and Lange, 2016, and S1–S3, Chs. 4–5), so our results can be straightforwardly compared with previous findings.

Task The task is a version of the well-known wug test (Berko, 1958, discussed in §2.2.1): participants were given a novel noun, such as *wug* (or in our case the more Germanic *Vag*), and asked to produce its plural inflected form. Our experiment includes an additional dimension. Along with the plural form, participants were asked to indicate the presumed grammatical gender of the noun by selecting the corresponding article for its singular form (Fig. 6.2).

We had two motivations for adding the gender task. Firstly, as earlier wug test studies have found weak to absent effects of gender on German plural inflection (§3.3), we sought an experimental design which would compel participants to attend to the gender of the noun. Secondly, we wanted participants to generate the full joint distribution over grammatical gender (G) and inflection class (C), so that we could evaluate their regularization behavior with respect to all three strategies identified by Ferdinand et al.

Procedure After providing consent, participants completed an onboarding task, in which they had to provide the gender and plural form for 12 real German nouns. Participants had to answer these questions correctly to proceed to the experiment. After the onboarding, participants were randomly assigned to one of three lists counterbalanced for presentation order of gender (e.g. "Der/Die/Das Vag" v.s. "Das/Der/Die Vag"). Within each list, the 24 test items were presented in randomized order. We publicly release the data.[4]

Participants We recruited 120 speakers with German as a first language to complete an online survey using the platform Prolific.[5] Speakers were compensated at the rate-adjusted equivalent of 11.50 USD per hour. Participants in S1–S3 were ineligible for this study.

Analysis Following Ferdinand et al. (2019), we quantify the entropy in the distribution produced per participant, and use it to classify participant behavior. Ferdinand et al. assume

---

[4] `https://github.com/kmccurdy/german-wug-data/`

[5] `https://www.prolific.co`

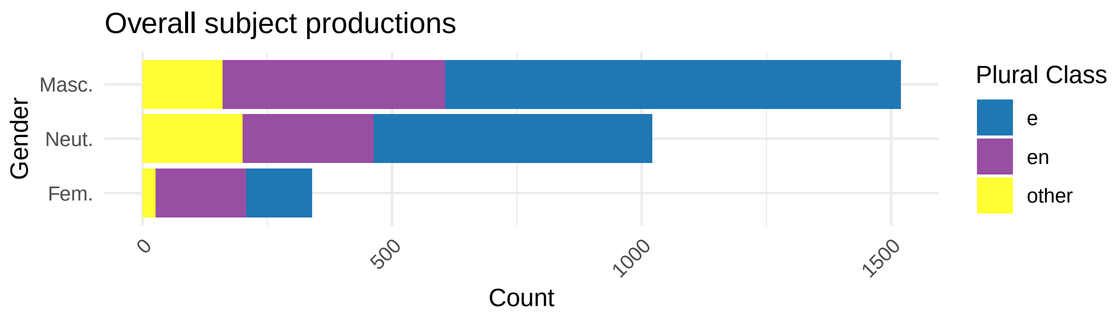


Figure 6.3: Gender and plural productions from participants. Compare to reference distributions in Fig. 6.1.

that participants with entropy measures within the 95% confidence interval (CI) bounds show behavior consistent with *probability-matching* the relevant distribution. To define probability-matching behavior, we simulate experimental draws over 24 items by sampling from the relevant joint categorical distributions. For each reference distribution, we first sample 10^5 grammatical gender assignments for the items, then plural class assignments conditional on the sampled gender. We calculate a more conservative 90% CI by taking the 5th and 95th percentiles of the resulting simulations. Participants with entropy measures below the 5% CI bound are classified as *regularizers*, and above the 95% are *variabilizers*, with respect to the same distribution. We build on Ferdinand et al.'s approach by also considering the *type* of regularization observed: overall reduction in variation (i.e. reducing $H(C)$) versus conditional reduction in variation (reducing $H(C \mid G)$, i.e. increasing $MI(C;G)$ or its normalized equivalent $U(C \mid G)$).

6.2.4 Results and Discussion

Fig. 6.3 presents the overall distribution of gender and plural productions from all participants (compare to the reference distributions in Fig. 6.1). We see considerable variation in gender and plural class assignment, which does not appear to be driven by strong item-level biases (c.f. Fig. 6.4).

Do speakers regularize overall variation? Fig. 6.5 shows the 90% CIs for the two reference distributions, and the observed range of speaker values, for our entropy-based measures. Overall variation is shown in the top row. For gender $H(G)$, most speakers' productions are consistent with probability-matching either reference distribution, falling within both CIs. For

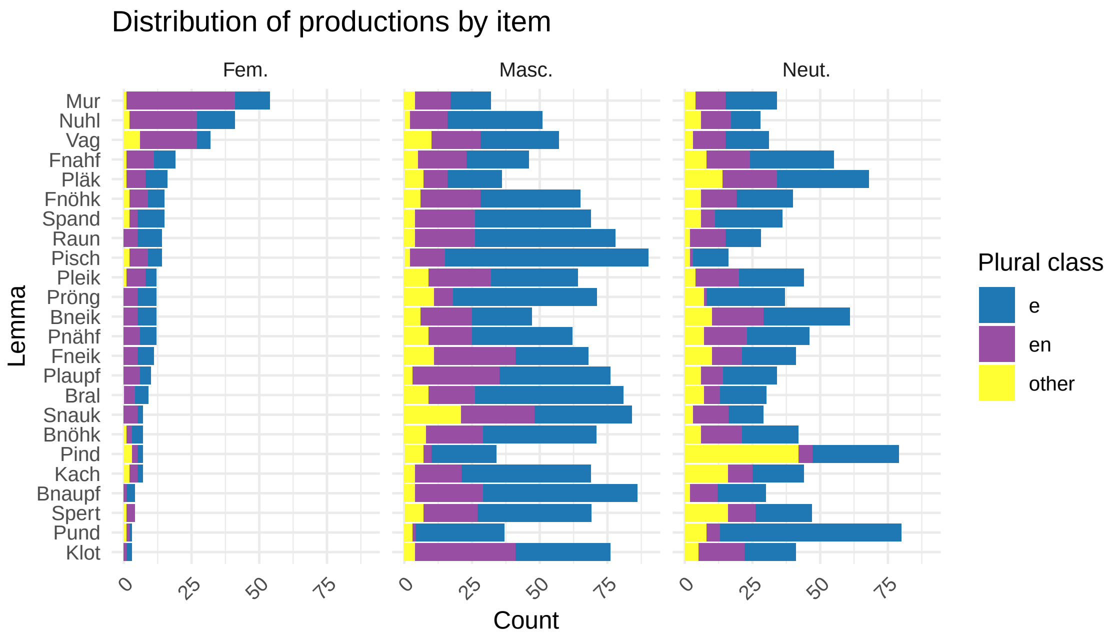


Figure 6.4: Gender and plural productions by item.

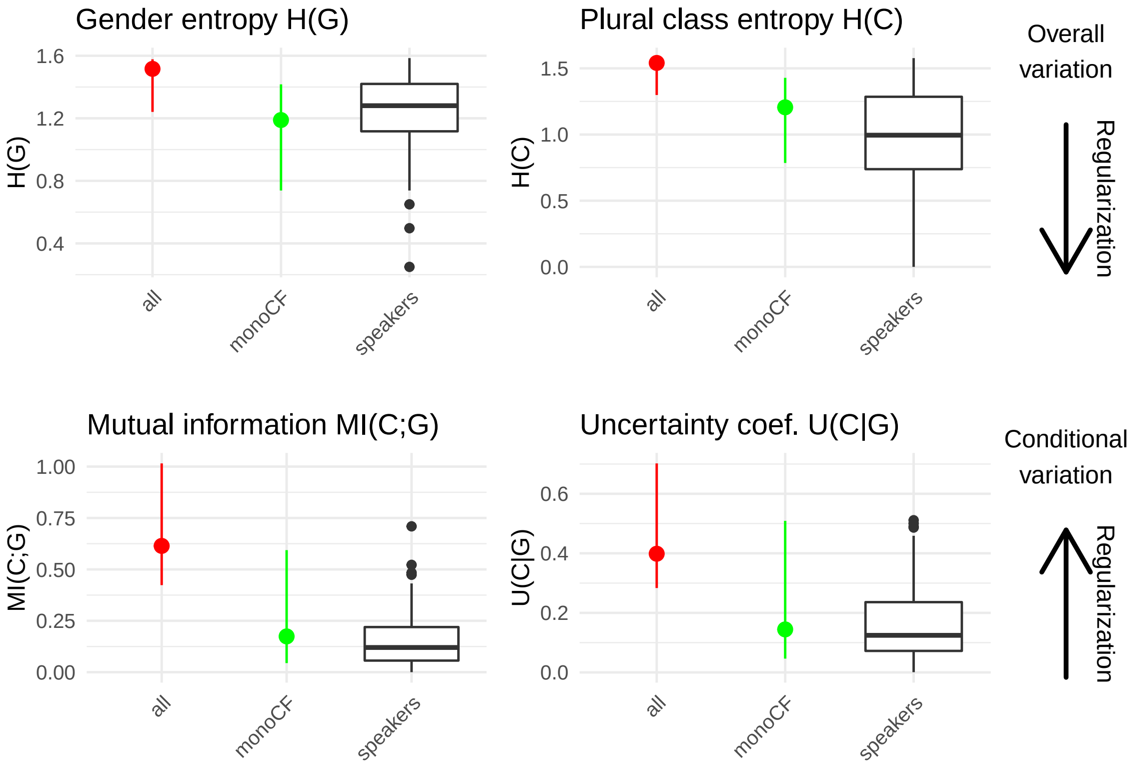


Figure 6.5: 90% CIs for reference distributions (all and monoCF nouns), and observed values for speakers. Speakers may regularize overall variation in $H(C)$, but do not appear to regularize $MI(C;G)$.

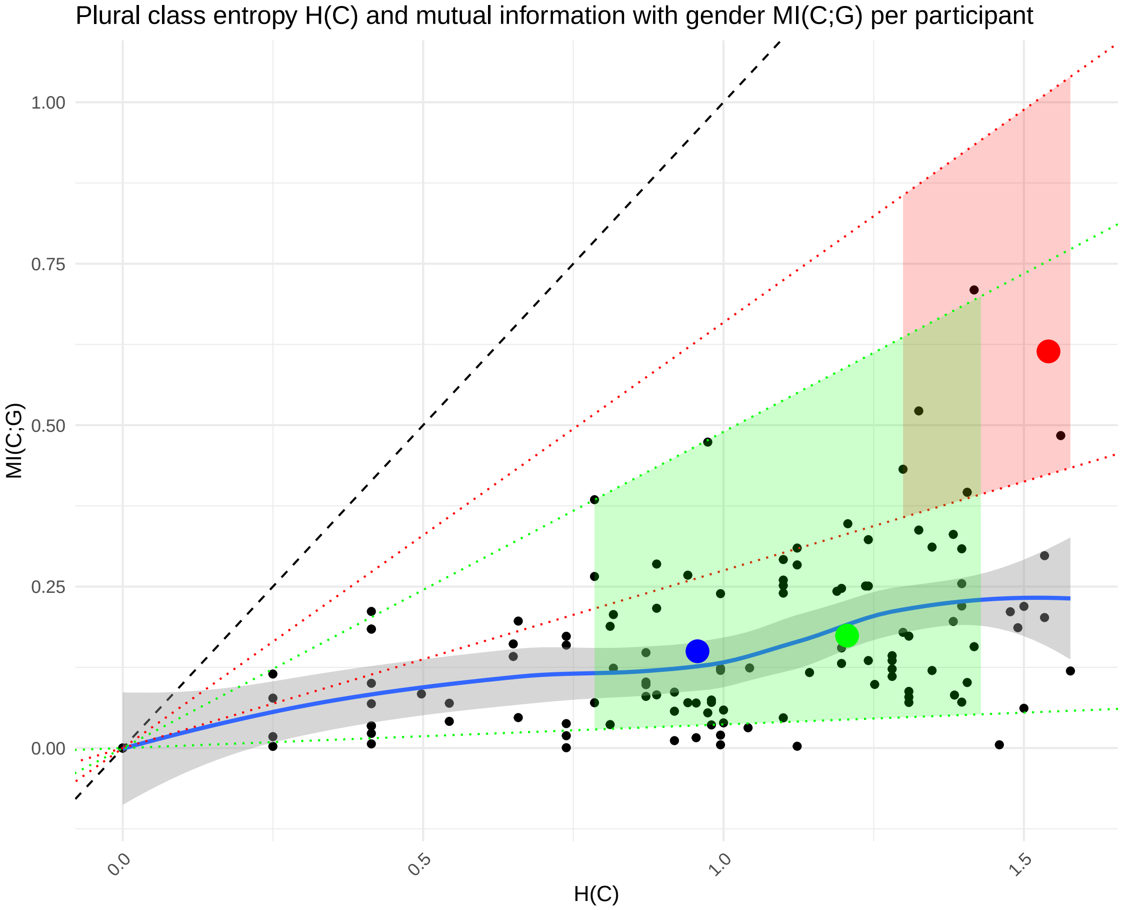


Figure 6.6: H(C) and MI(C;G) by participant. Color boxes and lines show 90% CI for all (red) and monoCF (green) nouns; color dots show reference values. The blue dot shows the speaker grand mean, and the blue line shows a Loess regression fit of speaker MI(C;G) on H(C). Most participants are in the green box, consistent with probability-matching the monoCF noun distribution.

plural class $H(C)$, we see some evidence for regularization: 75% of speakers reduce variation below the all-nouns 5% CI bound (c.f. Tab. 6.2). The bulk of those speakers show variation consistent with probability-matching the monoCF distribution, although 27% also fall below the 5% CI bound. In sum, we have two possible interpretations: either speakers are insensitive to the phonological properties of the stimuli and a large majority regularize plural class (i.e., relative to the lexicon as a whole); or speakers condition on phonology and are mainly probability-matching to a phonologically similar subset of the lexicon. However, the further analysis below suggests that speakers *are* sensitive to phonology, which makes the latter interpretation more plausible.

Do speakers regularize conditional variation? The lower row of Fig. 6.5 shows 90% CIs and the observed distribution for the conditional variation measures $MI(C;G)$ and $U(C \mid G)$, where higher values indicate greater predictability given the conditioning factor. Here we have clear evidence that speakers do *not* regularize by conditioning on grammatical gender; in fact, they seem to be probability-matching to the level of gender-conditioned predictability found in the monoCF nouns, which is substantially lower than that of the full lexicon. Speakers could, in principle, use the stronger relationship between gender and inflection class found in the full lexicon to make predictions about the stimuli, but they do not. This result is surprising given the importance of gender in both linguistic analyses (e.g. Augst, 1979; Wiese, 1996; Bittner, 1999) and recent models (e.g., recent neural network models make predictions that are consistent with the level of gender conditioning in the full lexicon; Goebel and Indefrey, 2000; McCurdy, Lopez and Goldwater, 2020*a*; Dankers et al., 2021). Our information-theoretic analysis suggests that speakers in fact condition on phonological form at the expense of predictability due to gender.

Interestingly, this reduced level of gender conditioning $MI(C;G)$ appears consistent relative to plural variation $H(C)$, although it need not be: speakers who vary plural class more could in principle introduce more gender conditioning. Fig. 6.6 shows, for each individual participant, how much variation they produced over plural class $H(C)$ — farther right on the x-axis indicates a more varied set of plural classes — and how much that variation was influenced by grammatical gender $MI(C;G)$ — higher on the y-axis indicates more gender-conditioning, i.e. a tighter statistical coupling between gender and plural class. The dotted black line shows $MI(C;G) = H(C)$, the theoretical maximum statistical coupling: a point on that line would represent a speaker whose plural class assignments were fully explained by grammatical gender, for example always assigning masculine nouns to the *-e* plural class and feminine nouns to the *-en* class. We see that even speakers who produce as much plural class variation as observed in the lexicon ($H(C) > 1.3$) are mostly below the red box, meaning their gender-conditioning $MI(C;G)$ is more typical of the monoCF distribution.

6.2.5 General Discussion

Our findings demonstrate that the regularization/probability-matching framework developed in the artificial language learning literature can also describe behavior in natural language tasks. Our work is not the first to show this; Hendricks et al. (2018) used this framework to

All nouns	Var. $H(C)$	Prob.-match	Regl.
Variabilize $U(C\|G)$	1% (1)	21% (25)	**60% (72)**
Probability-match		3% (4)	10% (12)
Regularize			
N/A $(H(C) = 0)$			5% (6)
monoCF nouns	**Var. $H(C)$**	**Prob.-match**	**Regl.**
Variabilize $U(C\|G)$	2% (2)	8% (10)	3% (4)
Probability-match	6% (7)	**58% (69)**	18% (21)
Regularize			1% (1)
N/A $(H(C) = 0)$			5% (6)

Table 6.2: Speaker strategy classification. Relative to all nouns, most speakers regularize overall plural class variation while *increasing* variability with respect to gender (upper table). Relative to monoCF nouns, most speakers probability-match overall and gender-conditioned plural class variation (lower).

study variable gender assignment in a Germanic dialect, finding that some children regularized while others probability-matched the variation in the adult distribution. To the best of our knowledge, however, we are the first to use lexical statistics as a reference distribution to evaluate regularization behavior in a natural language psycholinguistic experiment.

We suspect that probability-matching lexical statistics provides a stronger account for our results than most formal models. The substantial variation within items (c.f. Fig. 6.4) suggests a fundamental incompatibility with any models that make strong item-level predictions, which would encompass most rule-based models (e.g. Mugdan, 1977; Yang, 2016). Exemplar-based models (e.g. Hahn and Nakisa, 2000) may better handle such variability, but doing so appears to require extensive fine-tuning (c.f. Rosen, 2022). As noted earlier, parts of the German plural system are readily described by rules — our findings apply to the subset of the lexicon which shows less predictability. That said, many linguistic accounts of German inflection have proposed high-level rules based on grammatical gender (e.g. Augst, 1979; Wiese, 1996; Bittner, 1999), and neural models of German inflection learn behavior consistent with such rules (Goebel and Indefrey, 2000; McCurdy, Lopez and Goldwater, 2020*a*; Dankers et al., 2021). Our findings challenge such accounts: speakers *could* regularize by conditioning on gender to the extent observed in the German lexicon as a whole (i.e. 40% of plural class variation, c.f. Tab. 6.1), but instead they match the lower level of gender conditioning typical of the phonological class (15-16%). This accords with other linguistic accounts which consider gender subordinate to phonology (e.g. Mugdan, 1977; Spreng, 2004). Furthermore, our study's experimental design explicitly foregrounds gender by forcing participants to select both the article and plural class for each noun. This means that our results likely represent a *ceiling* for gender conditioning on these stimuli. Previous studies with the same stimuli have presented the article instead, and found weaker or absent effects of gender (Marcus et al., 1995; Zaretsky and Lange, 2016; McCurdy, Lopez and Goldwater, 2020*a*).

Conditional variation seems to play a paradoxical role in these results. On the one hand, there is strong evidence for phonological conditioning: speaker behavior is consistent with the lexical statistics of a phonologically similar subset of the lexicon, rather than the lexicon as a whole. On the other hand, we have two mysteries. Firstly, this phonological conditioning only appears at the level of word class; phonology does not seem to drive strong biases for individual items. Secondly, this phonological conditioning comes at the expense of gender conditioning: participants make grammatical gender *less* informative than it is in the lexicon as a whole. It is unclear how these trends relate to artificial language learning studies, which have found that adult learners tend to condition on lexical identity (i.e. reducing variation across nouns by assigning each noun to one lexical class; Smith and Wonnacott, 2010; Samara et al., 2017). Johnson et al. (2020) find that high mutual information (i.e. low i-complexity; Ackerman and Malouf, 2013) benefits learning for neural networks, but not for speakers, while low overall entropy (i.e. low e-complexity) benefits both. Our results echo their findings, as speakers appear to reduce overall entropy ($H(C)$), but unlike neural models, do not increase mutual information ($MI(C;G)$).

6.2.6 Conclusion

In this work, we take an information-theoretic measure of regularization developed for artificial language learning research, and use it to analyze experimental results in the natural-language domain of German plural inflection. We consider two possible points of reference — the lexicon of German nouns as a whole, and a restricted subset with a particular phonological shape — and find that speaker behavior is best described as probability-matching the lexical statistics of the latter phonologically-conditioned distribution. Although speakers could plausibly regularize by conditioning on grammatical gender (as predicted by the statistics of the overall lexicon), instead they appear to probability-match the lower level of gender conditioning seen on phonologically similar nouns. We demonstrate that lexical statistics can predict how

Study	$H(C)$	$H(C \mid G)$	$MI(C;G)$
S2	1.928	1.920	0.008
S3	1.930	1.918	0.013
S4	1.556	1.521	0.035

Table 6.3: Table 5.5 updated to include S4. Overall mutual information between gender and plural class increases in S4, but remains numerically small.

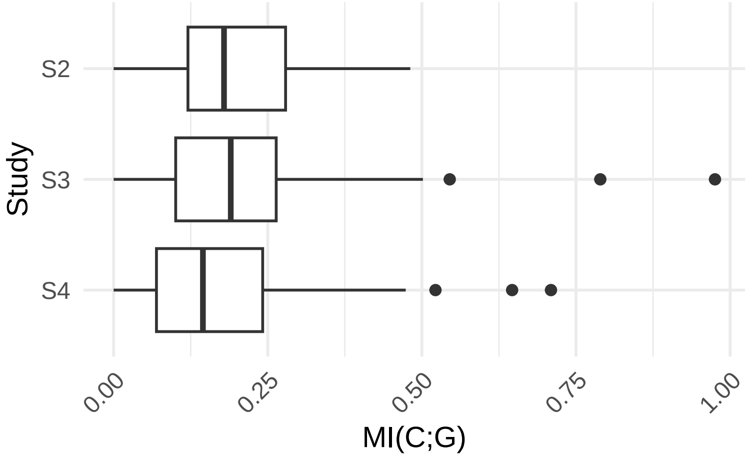


Figure 6.7: Distribution of mutual information (MI) between plural class and grammatical gender per participant per study. Observed values are similar across studies, with participants in S4 tending toward slightly *lower* gender conditioning on average.

speakers generalize lexical attributes to novel items, connecting artificial language findings with natural language behavior.

6.3 Discussion

The analysis presented above indicates that most speakers appear to probability-match to the distribution over plural classes for monosyllabic nouns in the German lexicon, indicating syllable structure as a key factor influencing generalization — more important than the statistically robust cue of grammatical gender. Why might this be the case? It's difficult to say precisely, given that syllabicity is confounded with many other important lexical properties. The biggest one is likely word length: monosyllabic nouns are typically shorter than nouns with multiple syllables. Famously, following Zipf's law, shorter words are also more frequent (e.g. Bentz and Ferrer Cancho, 2016); moreover, shorter and more frequent words typically have many other cognitively-relevant properties, such as higher concreteness or imageability (e.g. Reilly and Kean, 2007) and earlier age of acquisition (e.g. Brown, 1973). Although these properties are crucially dependent upon context missing from the experiments considered in this dissertation — the novel stimuli have no token frequency by design, are presumably encountered by participants for the first time in the experimental context, and have no concrete or abstract semantic referent — it's possible that their strong association in the extant lexicon has some effect on generalization in these studies. Nonetheless, I focus on monosyllabicity as a property which inheres in the word form alone, and therefore must be accessible to study participants. While word length is also apparent from the written string, the stimuli here also vary in their length; monosyllabicity alone is the property which unites them. Nevertheless, it is possible that monosyllabicity is not the crucial property informing generalization, but it is confounded with some other property of interest. In Chapter 7, I consider a further phonologically-informed reference distribution; otherwise, I leave the question open for future work.

Here, I briefly compare these results to the findings in Chapter 5. Did the joint production task in S4 yield a stronger statistical relationship between grammatical gender and plural class?

Paradoxically, the answer appears to be *yes (very slightly) overall*, but *no at the speaker level*. Table 6.3 shows relevant measurements for the aggregated results from studies S2, S3, and S4. Of the three studies, S4 shows the highest mutual information (MI) between grammatical gender and plural class, though this value remains far below the observed lexical values reported in Table 2 of §6.2. So does this mean that speakers show more gender conditioning on average in S4? Figure 6.7 indicates that the answer is no — individual speakers have roughly equivalent, if not slightly lower, rates of gender conditioning in S4 compared to S2 and S3. This paradox reflects the skewed distribution of gender classes in S4 (see §6.2, Fig. 3), where S2 and S3 had equally balanced gender presentation by design. The key signature of gender conditioning is greater use of *-en* plurals for feminine nouns. Participants in S4 do show this pattern to a greater extent than in S2 and S3, but they also only rarely assign feminine gender to these stimuli. This results in slightly higher mutual information between gender and plural class *overall*, but each individual participant is less likely to encounter (i.e. assign) a feminine noun, and thus show slightly lower levels of gender conditioning on *average*. In any case, as depicted in Figure 6.7, individual participants show very similar levels of gender conditioning across all three studies, and these levels are compatible with probability-matching to the joint distribution of grammatical gender and plural class observed for monosyllabic nouns in the German lexicon.

6.4 Conclusion

This chapter presents findings from S4, the fourth behavioral experiment in this dissertation using the novel noun stimuli developed by Marcus et al. (1995). In this study, German speaking participants were tasked with producing the joint distribution over grammatical gender and plural inflected forms for the stimuli. Participants showed similar levels of gender conditioning as seen in the previous experiments S2 and S3. This is incompatible with the statistical relationship observed in the overall lexicon, but compatible with the interpretation that participants generalize by probability-matching to a phonologically-conditioned lexical distribution — represented by monosyllabic nouns in the analysis in §6.2. This result suggests first– and second-order levels of conditioning: speakers probability-match to a distribution primarily conditioned on phonology (i.e. syllabic shape), and secondarily on grammatical gender. In Chapter 7, I evaluate whether this hypothesis — that speakers probability-match to a phonologically conditioned lexical distribution — adequately describes the behavioral data from all studies.

Chapter 7

Overview and Synthesis of Results

This dissertation investigates the question of German plural generalization, using recent distribution-focused methodological developments in behavioral and computational research to revisit this longstanding problem in linguistics and cognitive science. In the literature review of Chapter 2, I build on the framework of Herce (2019) to present a novel analysis of three distinct theoretical linguistic views of morphological regularity — rule generation, type frequency, and predictability — and how these perspectives broadly relate to morphological generalization in two domains: the observed behavior of speakers in psycholinguistic experiments, and modeling assumptions in computational experiments. Chapter 3 reviews the literature on German plural generalization through the lens of this tripartite conceptual organization. Chapters 4, 5, and 6 present a series of behavioral and modeling experiments on plural generalization using the novel German noun stimuli developed by Marcus et al. (1995).

This overview chapter synthesizes the behavioral and modeling results presented in previous chapters. I combine the production data from all wug tests to evaluate all of the models considered in this dissertation. The analysis in previous chapters points toward additional relevant comparisons, both in terms of models and of relevant baselines. For this reason, I introduce new models and baselines (§7.1), and evaluate their predictions along with previously described models in an overall comparison (§7.2). Based on the results (§7.3), I conclude that German speakers generalize plural inflection by probability-matching to a phonologically-conditioned lexical distribution, and discuss the implications for theories of morphological regularity (§7.4).

7.1 Additional Models of the German Noun Lexicon

The results for these models are presented within the overall comparison in the next section. Note that the neural and symbolic models described in previous chapters were evaluated by sampling the single most likely classification for each item, and therefore required sampling over variable training regimes — over different random seeds in the case of the neural model,[1] and over different dataset samples in the case of the symbolic model. In contrast, the three models considered here predict a probability *distribution* for each evaluation item. This means that, rather than sampling from these models, we can compare their predicted distributions directly to the observed distributions produced by speakers.

7.1.1 Bayesian Classifier and Naive Bayes

As discussed in §2.3.2, Bayesian models are particularly well-suited to capture probability-matching behavior due to their tendency to converge on the posterior distribution. This suggests that, within the lexical modeling regime used in this dissertation — namely, learning a mapping from a input lexeme to an output inflected form — Bayesian modeling may best express the type frequency view of regularity, and its associated probability-matching behavior. The analysis in Chapter 6 indicates that speaker behavior is likely best understood as probability-matching to a lexical distribution conditioned primarily on prosody (i.e. monosyllabicity), and perhaps secondarily on grammatical gender. The neural and symbolic modeling predictions presented in Chapters 4 and 5 do not adequately capture this variability in speaker productions; neural models in particular condition heavily on statistical cues such as grammatical gender, leading to overconfident predictions. Perhaps the distributional focus of Bayesian learners lets them better model the speaker variability we observe in this domain.

Method As in the rest of the dissertation, the modeling task is to learn a mapping from an input noun in the singular form, to an output plural inflected form. The model is trained on the German noun lexicon and evaluated by comparison to speaker behavior. The Bayesian model, however, differs from the other models in several key respects.

The most critical difference is in the problem formulation and representation. Due to

---

[1] See also the discussion of lower-ranked degenerate sequences within beam search for the neural model, McCurdy, Goldwater and Lopez (2020), Chapter 4.

the complexity of sequence modeling in a Bayesian framework (c.f. §2.3.2), I significantly simplify the modeling task in the interest of an expedient preliminary investigation. Instead of generating a sequence of characters for the output form, I use a Bayesian logistic regression model which classifies an input noun as belonging to one of six plural classes. I also remove sequence modeling from the input representation. Instead of a character sequence, each input noun is represented by three features: grammatical gender and final character (which account for a substantial amount of plural class variation in the lexicon; §3.1), and number of syllables (to capture prosodic conditioning). Even this simplified feature representation proves difficult to fit, due to the categorical variable of final character, which can take one of 26 unique values. Preliminary experiments showed that this categorical variable can be easily replaced with a three-dimensional continuous representation. For each character, I retrieve the first two dimensions of its pretrained German FastText embedding (Bojanowski et al., 2017), and use these two values and their interaction to represent the character in the Bayesian model. This modification substantially stabilized and sped up training, with little to no impact on the model's predictive accuracy. Although these simplifications drastically alter the modeling task, they let us answer the key exploratory question driving this analysis, namely whether a Bayesian model of the lexicon can capture the probability-matching behavior of speakers.

In keeping with the exploratory nature of this analysis, the Bayesian model also differs from the earlier models in terms of its training data and regime. I train the model on the CELEX lexicon (Baayen et al., 1995) rather than UniMorph. An additional difference is that the model is trained on the entire lexicon; I do not hold out a validation set (no hyperparameter selection), or a test set (primary comparison is not accuracy, but speaker generalization to novel words).

Bayesian Classifier I specify a categorical regression model with the multivariate logit link function (i.e. softmax) and predictor variables x_S (syllable count), x_{F1} and x_{F2} (2D representation of final character), and x_G (grammatical gender), with a weakly informative Student's T-distribution prior for each coefficient. The probability p_k for a specific plural class k is given as follows:

$$log(\frac{p_k}{1-p_k}) = \beta_0 + \beta_1 x_S + \beta_2 x_{F1} + \beta_3 x_{F2} + \beta_4 x_{F1} x_{F2} + \beta_5 x_G \tag{7.1}$$

$$\beta \sim StudentT(0, 2.5, 3) \tag{7.2}$$

I train two variants of the model, one as above, and one without grammatical gender as a predictor. Both models are fit in STAN (Carpenter et al., 2017) using the `brms` library (Bürkner, 2017, 2020) in R (R Core Team, 2023). Both models successfully converged.

Naive Bayes In addition to the Bayesian classifier, I also implement a Naive Bayes classifier using the same feature representations and predictors. Categorical probabilities for grammatical gender, number of syllables, and prior probability of each plural class are computed directly from observed values in the lexicon. For the continuously-valued final character feature, I fit a two-dimensional multivariate Gaussian using the `mvtnorm` library in R (Genz and Bretz, 2009). As in the case of other models, I assess two variants of the Naive Bayes classifier, one with grammatical gender as a predictor, and one without.

7.1.2 Nearest Neighbors

As discussed in §2.3.3 and §3.4, some researchers have argued that exemplar models are best suited to capture how speakers generalize morphology; for instance, Ambridge (2020) reviews an extensive body of literature and claims that exemplar models mirror how speakers generalize language over a range of domains. Milin et al. (2011) find that analogical generalization from lexical nearest-neighbors predicts how Serbian speakers inflect novel nouns in a wug task, and Blevins et al. (2017) build on this technique to model relationships between German noun paradigm cells. I include their exemplar model of nearest-neighbor classification here for comparison.

Method Again using CELEX, I follow the procedure described by Milin et al. (2011) and Blevins et al. (2017) to generate nearest-neighbor predictions for the novel noun stimuli developed by Marcus et al. (1995). For each stimulus item, I identify their lexical neighbors as all nouns with a normalized Levenshtein distance of less than $.5$, and take the resulting distribution over neighbors' plural classes as the prediction.[2] As for other models, I generate predictions both with and without grammatical gender.

---

[2] Note that this simple method lacks any smoothing mechanism, which makes it brittle; it does not generate any prediction in cases where no words meet the similarity threshold, i.e. no neighbors are found.

		e	en	s	er	zero	other
Uniform	-	0.20	0.20	0.20	0.20	0.20	0.00
Type freq.	UM	0.35	0.37	0.04	0.03	0.19	0.02
	CLX	0.27	0.48	0.04	0.04	0.17	0.00
Token freq.	CLX	0.21	0.45	0.02	0.03	0.29	0.00
Monosyllables	UM	0.66	0.14	0.13	0.05	0.01	0.01
	CLX	0.63	0.12	0.14	0.06	0.05	0.01
M95 Rhymes	UM	0.72	0.14	0.03	0.07	0.01	0.03
	CLX	0.67	0.15	0.04	0.12	0.00	0.03

Table 7.1: Calculated values for lexical baselines.

7.1.3 Additional Lexical Baselines

I include several relevant baselines in this comparison, based on the findings in previous chapters. All baselines are static: the same probability distribution is compared to each stimulus item. No baseline reported here incorporates grammatical gender. Table 7.1 gives the calculated values for each baseline.

Uniform I consider a baseline of uniform probability distribution over five plural classes, with zero probability assigned to the “other” category.

Type and Token Frequency I calculate general type frequency baselines from Unimorph and CELEX. As CELEX includes token frequencies, I include a token frequency baseline from this resource (Sonnenstuhl and Huth, 2002).

Phonologically Conditioned In light of the analysis in Chapter 6, I include two baselines covering only a restricted set of the lexicon, namely nouns which are phonologically similar to the stimuli developed by Marcus et al. (1995, M95). One baseline represents the plural class distribution over **monosyllabic** nouns, while the other includes only nouns which **rhyme with the M95 stimuli**. In both cases, nouns matching this criterion are selected, and the static distribution of plural classes is calculated from this lexical subset. I calculate each of these measures for both the UniMorph (UM) and CELEX (CLX) lexicon, resulting in four phonologically-conditioned baselines.

Study	N Participants	Chapter	Gender	Gender Counterbalanced?	N Samples
S1	150	4	N	-	24
S2	92	5	N/M/F	yes	72
S3	100	5	N/M/F	yes	72
S4	120	6	N/M/F	no	72
Total					240

Table 7.2: Summary of evaluation data from behavioral experiments.

7.2 Overall Model Comparison

This section evaluates all of the models used in this dissertation on all of the behavioral data collected. The goal of this analysis is to identify, for each stimulus item, the *distribution* of plural classes produced by speakers — and to evaluate, for each model, how closely their predictions on that item match the distribution produced by speakers. I use two information-theoretic measures — entropy and Kullback-Leibler (KL) divergence — to assess which model most closely approximates observed speaker behavior.

Behavioral Experiments Each model's predictions are evaluated on the distribution of speaker plural productions for the four behavioral experiments described in preceding chapters. The first study, **S1**, is described in Chapter 4. In this case, speakers were only exposed to one grammatical gender (neuter) with each item, so this study yields 24 observations (one per item). Studies **S2** and **S3** are described in Chapter 5. These two studies largely share the same experimental protocol; in both of these experiments, grammatical gender presentation was evenly counterbalanced for each stimulus item. The only difference is that participants in **S3** were offered additional remuneration for each item where they produced the same plural form as the majority of other participants. As discussed in Chapter 5, this intervention had almost no effect on participant generalization behavior. Study **S4** is described in Chapter 6. In this experiment, participants selected both the grammatical gender of the singular noun item and its plural inflected form, resulting in a highly skewed distribution of grammatical gender across items. In any case, I consider each grammatical gender and stimulus item combination separately, which results in 72 observations (24 items times 3 gender categories) from each of S2, S3, and S4. In total, then, I have 10 distributional samples (3 masculine, 3 feminine, and 4 neuter) of speakers' plural class productions for each of the 24 stimulus items, yielding 240 items as summarized in Table 7.2.

Family	Model	Description	Data	Gender?	N Samples
Symbolic	ATP	§2.3.1	UM	Gender	25
				-	25
Neural	RNN Encoder-Decoder	§4.2	UM	Gender	25
		§4.3.1		-	25
Exemplar	Nearest Neighbors	§7.1.2	CLX	Gender	Dist.
				-	
Bayesian	Classifier	§7.1.1	CLX	Gender	Dist.
				-	
	Naive Bayes	§7.1.1	CLX	Gender	Dist.
				-	
Baseline	Uniform	§7.1.3	-		Dist.
	Type Frequency		UM		
			CLX	-	
	Token Frequency		CLX	-	
	Phon - Monosyllables		UM	-	
			CLX	-	
	Phon - M95 Rhymes		UM	-	
			CLX	-	

Table 7.3: Summary of models and baselines compared. Baselines are static, while model predictions typically vary for each stimulus item. Each model is trained on either the UniMorph (UM) or Celex (CLX) lexical resource. See text for further description.

Models Table 7.3 summarizes the models and baselines compared in this analysis. For models which produce point estimates (i.e. single most likely classifications) rather than distributions, the number of samples is reported; for models and baselines which predict a probability distribution over plural classes, the predicted distribution is directly compared to the observed distribution of participant responses over each item. Each model is trained with and without grammatical gender.

Measures I use two information-theoretic measures to assess the fit between model predictions and speaker productions. The first, **Jensen-Shannon (JS) divergence** (Lin, 1991), calculates the degree to which a model's predicted distribution Q *diverges* from the reference distribution P produced by speakers. This measure takes into account the *contents* of the predicted distribution. For example, if a model assigns a probability of 70% to the plural class -*en* for some specific stimulus item, but 70% of speakers produced -*e* for that item, the JS divergence between these two distributions would be high.

Jensen-Shannon divergence is a smoothed, symmetric version of Kullback-Leibler (KL) divergence, or relative entropy. For a discrete reference distribution P and predicted distribution

Q, KL divergence is defined by the following equation:

$$D_{\mathsf{KL}}(P\|Q) = \sum_i P(i)\log\left(\frac{P(i)}{Q(i)}\right) \tag{7.3}$$

Importantly, KL divergence is *asymmetric*: $D_{\mathsf{KL}}(P\|Q) \neq D_{\mathsf{KL}}(Q\|P)$, and the two measures (known as *forward* and *reverse* KL divergence) penalize different types of mismatch between the two distributions. These differences can be useful for certain applications; however, for the present analysis, we're interested in any kind of divergence between the speaker and model distributions, regardless of direction. For this reason, I use the Jensen-Shannon divergence, which equally weights forward and reverse KL divergence from the mixture of P and Q:

$$M = \frac{1}{2}(P+Q) \tag{7.4}$$

$$JSD(P,Q) = \frac{1}{2}D_{\mathsf{KL}}(P\|M) + \frac{1}{2}D_{\mathsf{KL}}(Q\|M) \tag{7.5}$$

$$= \frac{1}{2}\sum_i P(i)\log\left(\frac{P(i)}{M(i)}\right) + \frac{1}{2}\sum_i Q(i)\log\left(\frac{Q(i)}{M(i)}\right) \tag{7.6}$$

JS divergence can be interpreted as the mutual information (Eq. 3.1) between the mixture distribution M and an equiprobable indicator variable Z which separates P and Q, i.e. sets i to $P(i)$ when $Z = 0$ and $Q(i)$ when $Z = 1$. This has two relevant implications: 1) JSD is bounded between 0 and 1, as $H(Z) = 1$; and 2) JSD can be expressed in bits.

The second measure is simply the **entropy** (Eq. 2.1, §2.1.2) of the observed distribution. As discussed in §2.2, entropy has been used to measure *probability-matching* behavior in the experimental literature on regularization: participants who produce a distribution with roughly the same entropy as their training inputs can be classified as probability-matching, while participants who produce lower-entropy distributions are said to regularize (c.f. Ferdinand et al., 2019). In Chapter 4, we saw that both neural and symbolic models tended to produce more concentrated predictive distributions relative to speakers, and in Chapter 6, we saw that speakers tend to probability-match to certain lexical distributions of plural classes. In this chapter, I compare the observed entropy for speaker productions for each stimulus item to the entropy of model predictions. Unlike KL divergence, this measure is not sensitive to the specific contents of the predicted distribution; I use it to assess whether model predictions are more or less variable compared to speaker productions.

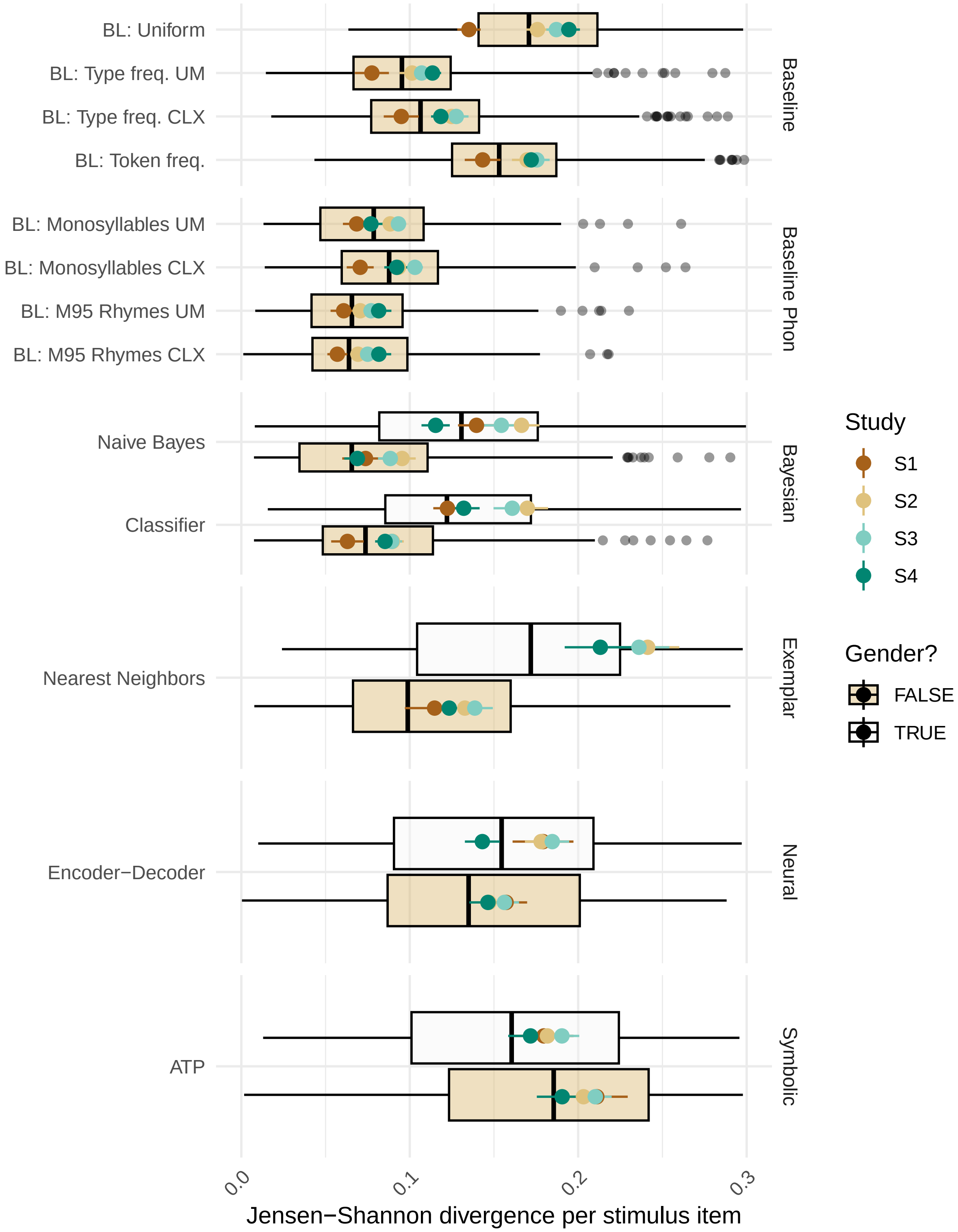


Figure 7.1: Overview: Jensen-Shannon divergence between speaker productions and model predictions for each stimulus item in each behavioral experiment. See model descriptions in Table 7.3. Boxplots summarize median values and interquartile range from items in all experiments, while the colored dots give separate mean values and standard error bars for each of the four behavioral experiments (Table 7.2). For ease of visualization, divergences greater than 0.3 bits are not plotted. A phonologically conditioned baseline (M95 Rhymes CLX; Table 7.1) shows the lowest median JSD, followed by the Naive Bayes classifier trained without gender.

			JSD		Entropy	
		Gender?	Mean (SD)	Median	Mean (SD)	Mean Diff.
Baseline	Token freq.	-	0.169 (0.06)	0.159	1.774 (0)	0.297
	Type freq. CLX	-	0.121 (0.06)	0.107	1.824	0.318
	Type freq. UM	-	0.104 (0.05)	0.095	1.966	0.414
	Uniform	-	0.181 (0.05)	0.171	2.322	0.743
	M95 Rhymes CLX	-	**0.073** (0.05)	**0.064**	1.5	0.276
	M95 Rhymes UM	-	0.075 (0.05)	0.066	1.377	0.324
	Monosyllables CLX	-	0.094 (0.05)	0.088	1.7	0.274
	Monosyllables UM	-	0.084 (0.05)	0.079	1.524	**0.270**
Bayesian	Classifier	Gender	0.151 (0.09)	0.129	1.388 (0.74)	0.597
	Classifier	-	0.086 (0.05)	0.074	1.831 (0.27)	0.404
	Naive Bayes	Gender	0.145 (0.08)	0.132	1.045 (0.44)	0.569
	Naive Bayes	-	0.083 (0.07)	0.066	1.493 (0.27)	0.317
Neural	Encoder-Decoder	Gender	0.17 (0.09)	0.162	0.491 (0.41)	1.092
	Encoder-Decoder	-	0.151 (0.08)	0.138	0.72 (0.45)	0.882
Exemplar	Nearest Neighbors	Gender	0.238 (0.15)	0.216	0.909 (0.63)	0.740
	Nearest Neighbors	-	0.13 (0.09)	0.104	1.423 (0.57)	0.554
Symbolic	ATP	Gender	0.181 (0.09)	0.171	0.512 (0.61)	1.084
	ATP	-	0.202 (0.1)	0.205	0.363 (0.54)	1.222
Speakers	S1				1.801 (0.22)	0.124
	S2				1.669 (0.33)	0.188
	S3				1.669 (0.26)	0.164
	S4				1.326 (0.29)	0.266

Table 7.4: Overview of results. For each baseline and model, the mean (standard deviation) and median Jensen-Shannon divergence (JSD) from speaker productions, mean item-level entropy, and mean difference to speaker production entropy for the same item. Speaker entropy values are included for comparison, with item-level differences calculated from the average across all studies.

7.3 Summary of Findings

Table 7.4 presents the full set of results. Figure 7.1 visualizes the Jensen-Shannon divergence (JSD) values, and Figure 7.2 visualizes the entropy measurements. Here, I highlight several key findings.

Phonologically conditioned lexical baselines are more speaker-like than nearly all models. Across all experiments, the M95 Rhymes CLX baseline has the lowest average (mean) JSD from the item-level plural class distributions produced by speakers. This static baseline diverges by only 0.073 bits on average from *every* item distribution produced by speakers, narrowly edging out the same baseline calculated on the UniMorph corpus (0.075 bits). Rounding out the top 5 on this measure are the Naive Bayes classifier trained without gender (0.083 bits average divergence), the Monosyllables UM baseline (0.084), and the Bayesian Classifier trained without gender (0.086). The two rhyme baselines also take best fits when measuring *median* JSD, with M95 Rhymes CLX coming in first (0.064 bits), and M95 Rhymes UM tied with gender-free Naive Bayes for second place (0.066); fourth place goes to the gender-free Bayesian classifier (0.74), and fifth to the Monosyllables UM baseline (0.079). Finally, in terms of variability, all phonological baselines show more speaker-like levels of entropy than all computational models (again with the exception of Naive Bayes). Note that the relative success of models trained without gender is unsurprising in light of the findings in Chapters 5 and 6; nonetheless, it may be surprising that such a diverse range of statistical learning models are consistently outperformed by static baselines in terms of matching to speaker behavior.

Bayesian models show the most speaker-like variability. Figure 7.2 plots item-level entropy measurements for the plural class distributions produced by speakers and predicted by models. Among the models trained without grammatical gender, the Bayesian Classifier, Naive Bayes, and Nearest Neighbor models generate predictions with similar entropy values to those produced by speakers; however, including grammatical gender tends to concentrate predictions and therefore lower entropy for nearly all model classes. This is especially notable for feminine gender nouns — given the strong statistical association between feminine gender and plural class in the German noun lexicon (§3.1), all models predict distributions with substantially lower entropy than the variable productions we observe from speakers in behavioral experiments. For nonfeminine nouns, however, the Bayesian Classifier predicts robustly variable distributions, well within the range of speaker productions, while the Naive Bayes and Nearest Neighbor models fall in a lower but intermediate range. As discussed in Chapter 4, we see that both the neural Encoder-Decoder and symbolic ATP models produce much more concentrated predictive distributions. To use the terminology of the regularization literature, we might say that the Encoder-Decoder and ATP are *regularizing* relative to speakers, while the Bayesian and Nearest-Neighbor models are more inclined to *probability-match* (cf. Ch. 6). Gender-free Bayesian Classifier and Naive Bayes appear to match both the variability (measured in entropy) and contents (JSD) of speaker productions better than other models — but in both respects, they still fall behind phonologically-conditioned lexical baselines (Table 7.4).

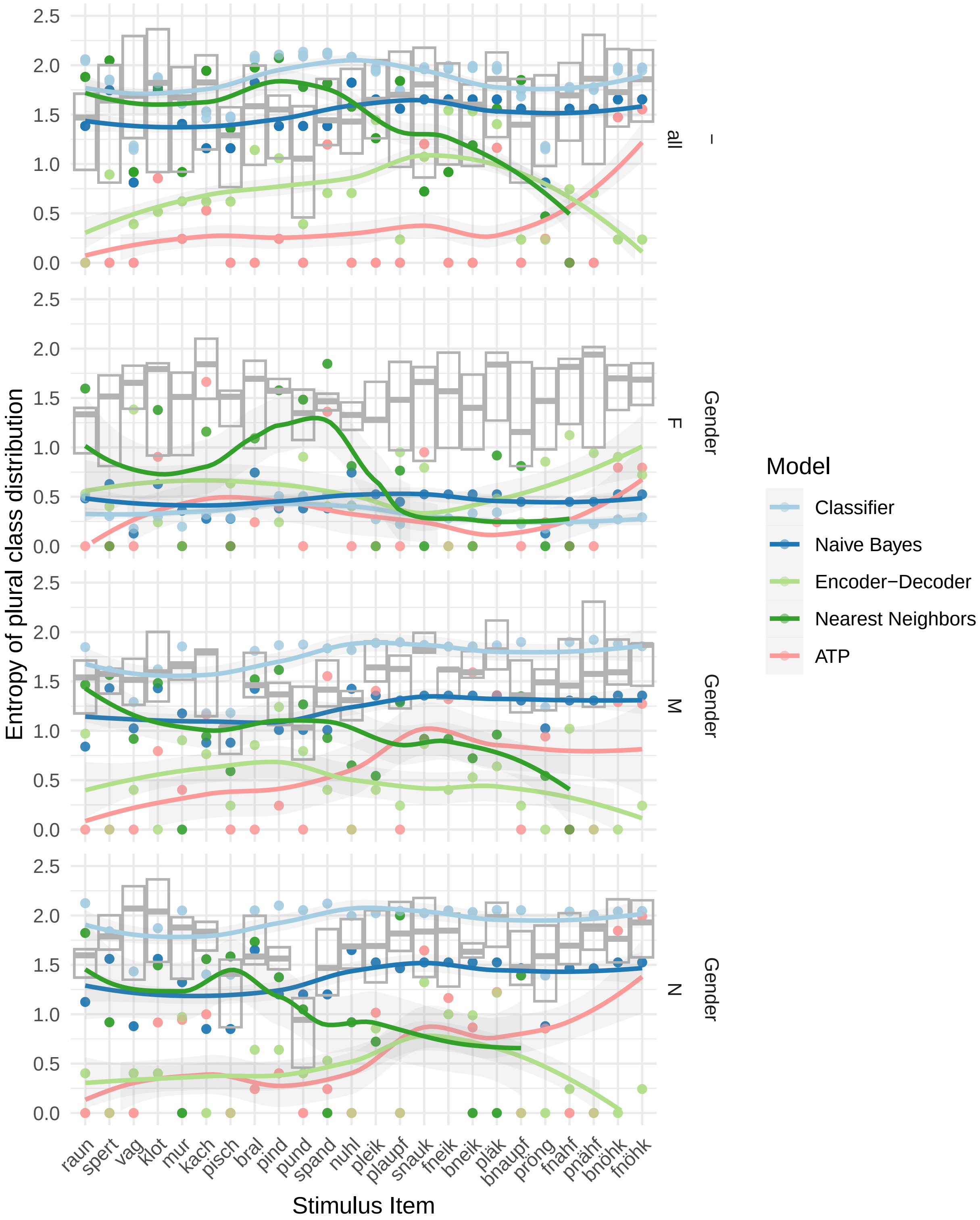


Figure 7.2: Entropy per stimulus item and grammatical gender. Measurements for model predictions are plotted in points and smoothed lines, for models trained without gender (top panel) and with gender; measurements for speaker productions are shown in grey crossbars (max, median, min), including all gender combinations in top panel. Stimuli are ordered from highest (*Raun*) to lowest (*Fnöhk*) likelihood (c.f. Table 3.6). Trend lines are generated by loess (local nonparametric) regression. Note that only Bayesian model predictions (blue lines) are consistently within the entropy range of speaker productions (grey boxes).

Figure 7.2 also highlights interesting differences in how model predictions interact with the plausibility of a given stimulus item, as measured by likelihood under a character n-gram model (c.f. Table 3.6). The Bayesian models, like the speakers across all experiments, maintain a consistent level of variability (measured by entropy) from the most likely novel noun (*Raun*, a Rhyme stimulus) to the least (*Fnöhk*, a Non-Rhyme stimulus). The symbolic ATP shows an intuitively sensible pattern: it is most confident for likely nouns like Raun, and produces increasingly uncertain (i.e. variable) predictive distributions for highly unlikely nouns like Fnöhk. The exemplar Nearest Neighbor model shows the opposite trend — more variable predictions for plausible nouns like Raun, which have many neighbors, and less variability for nouns like Pläk with few neighbors; it doesn't even generate predictions for nouns like Fnöhk, where no existing nouns meet the similarity threshold.[3] The neural Encoder-Decoder model doesn't show any particularly robust trend in this respect, although it appears to have a slight tendency toward confident predictions for both high and low probability stimuli (both Raun and Fnöhk), and more variable predictions for stimuli in the middle of the likelihood range.

7.4 Discussion

In this section, I consider how the results reported here relate to the analysis of regularity and morphological generalization developed in Chapter 2. To review, linguistic theories have posited three different criteria[4] of morphological regularity which may influence how inflection classes generalize to novel words: rule generation, type frequency, and predictability. The rule generation account has developed largely within the generative linguistic tradition. Under this analysis, morphological generalization is primarily determined by constraints on the input to individual transformation rules, which in turn assign a single inflection class to output forms (c.f. Table 2.5); thus, this framework focuses on *input* properties of inflection classes considered in *isolation*. The type frequency account instead characterizes morphological generalization in terms of the *distribution* of inflection classes among *output* inflected forms. Finally, the predictability account also focuses on *distributions*, but flexibly integrates information from both *inputs and outputs*. While I have separated these conceptual dimensions for clarity (c.f.

---

[3] This limitation could be addressed by relaxing similarity thresholds as needed to achieve a minimum number of lexical neighbors, but such work is beyond the scope of this dissertation

[4] A fourth criterion, concatenativity, is discussed in §2.1 but not investigated in this dissertation.

Table 2.1), linguistic and computational researchers have found diverse ways to combine and apply these three approaches in practice.

The German plural system has attracted the attention of linguists and cognitive scientists because the statistical properties of the German noun lexicon make these three criteria separable. In contrast to the English past tense suffix *-ed*, for instance, which is regular in all three aspects — unrestricted in the input, highly frequent, and highly predictable — German plural classes show a wider range of properties associated with morphological regularity, making it possible to separately analyze how these dimensions influence inflection generalization. The rare class *-s* shows the least evidence of input restrictions, indicated by its appearance in a broad range of linguistic environments ($H(F, G)$ in Table 3.3); the two suffixes *-e* and *-en* have the highest type frequency, although neither commands a statistical majority in the lexicon (Table 3.1); and *-en* in particular is highly predictable conditional on feminine grammatical gender ($H(G)$ in Table 3.3). If we think that linguistic structure causally affects how speakers generalize (as many researchers do; §2.2), then these distinctions predict different generalization behaviors respectively: regularization with the minority variant *-s*, probability-matching overall type frequency, and conditional probability-matching (or even regularization) based on predictive cues such as gender (see also discussion in §2.2.2 and §3.3.1).

The goal of this chapter, and this dissertation more broadly, is twofold. Firstly, I aim to determine which of these predictions most accurately characterizes the plural generalization behavior of German speakers. The second goal is computational: as different classes of computational models have inductive biases favoring different types of generalization (§2.3), I aim to determine which class of computational model most closely approximates speaker behavior after training on the German noun lexicon. Indeed, the comparative empirical analysis of the second research question crucially supports the first — careful computational evaluation can help us interpret the behavioral evidence, which historically has been fragmented and contradictory in this domain (§3.3.3).

Based on the analysis in this chapter, I conclude that German speakers generalize plural inflection by *probability-matching to a phonologically-conditioned lexical distribution*. This outcome reflects aspects of both the type frequency and predictability accounts: speakers generalize a distribution of output inflected forms which is partly constrained by the input, but do not appear to integrate all cues in the input which could minimize uncertainty — notably, they are largely insensitive to the highly predictive cue of grammatical gender. Computational analysis supports this interpretation: speaker data is best fit on average by a simple lexical baseline phonologically matched to the stimuli (Table 7.4), with Bayesian models coming in second as the only computational model class which consistently matches speaker-level variability on the level of individual items (Fig. 7.2). Including grammatical gender as a predictor, however, worsens the fit of the Bayesian models; they have the correct bias toward frequency-matching, but do not appear to share the inductive biases that lead speakers to condition on certain factors (e.g. phonological similarity) and largely disregard others (e.g. grammatical gender).

With respect to the dimensions of regularity discussed in §2.1.1, these results support an account of morphological generalization primarily driven by *distributions* over *output* inflected forms and only partially constrained by certain *input* conditions. It is not immediately clear how to identify and represent the relevant input features. Considering Table 7.4, the relative success of the monosyllabic baseline suggests a role for coarse-grained phonological features such as syllable structure. The even greater success of the rhyme baseline — along with the consistent increase in *-s* plurals for nouns ending in *-k*, discussed in Chapter 4 — speaks to an influence of fine-grained phonological features; however, if a static baseline can outperform item-level model predictions, this suggests speakers cannot be particularly sensitive to the fine-grained features of specific stimuli. The effect of grammatical gender is similarly vague. Gender shows a statistically significant influence on plural productions in Studies S2, S3, and S4 (c.f. Table 7.2); however, as discussed in Chapter 6, the size of this effect is drastically reduced relative to what we observe in the lexicon (Table 3.3).

The scientific challenge, then, is to properly characterize the lexical distribution which informs how speakers generalize morphological inflection. Given the competitive performance of the rhyme baseline, it is not clear that any of the models considered here truly succeed in this task. We can, however, rule out certain theoretical perspectives based on these findings. The rule generation view of regularity may be descriptively adequate for some analytical purposes, but does not appear to predict linguistic generalization behavior by speakers; the behavioral results reported here are simply not explicable in terms of input conditions for rule application, as developed within the generative tradition and advocated computationally e.g. by Veríssimo and Clahsen (2014) and Yang (2016). Similarly, the predictability view of regularity, exemplified in this dissertation by the Low Conditional Entropy Conjecture (Ackerman and Malouf, 2013, §2.1.2), posits that speakers may use all possible sources of information to reduce uncertainty in morphological generalization. The results in this chapter indicate that this assumption is too strong: speakers are insensitive to the highly informative cue of grammatical gender, a finding which echoes some other studies (e.g. Dawdy-Hesterberg, 2014, discussed in §2.2.1). Of the three main theoretical approaches to German plural inflection (§3.2), both the minority and gender-conditioned default analyses appear decisively ruled out. While it is unclear whether schema theory (§3.2.3; Köpcke, 1988; Bybee, 1995) expressly predicts the behavioral outcomes reported here, it is certainly the most compatible theoretical proposal, given its central focus on output distributions. Computationally, Bayesian learners are inclined to probability-match, and therefore appear most suited to model speaker-like generalization; however, none of the models considered here have fully captured the correct distribution. I consider other modeling possibilities in the concluding chapter.

7.5 Conclusion

In this chapter, I evaluate a range of computational models and baselines using behavioral speaker data from the four studies described in previous chapters, and consider the implications for linguistic theories of morphological generalization. I find that a phonologically-conditioned lexical distribution shows the lowest average divergence from speaker productions across all experiments, and conclude that German speakers generalize plural inflection by probability-matching to such a distribution, at least for the small set of noun stimuli evaluated here. These results support an account of morphological regularity in which generalization is primarily driven by distributions over output inflected forms, and only partially constrained by certain input conditions. This account integrates elements of the predictability and type frequency criteria of regularity, but appears wholly incompatible with the rule generation criterion of morphological generalization.

Chapter 8

Conclusion

> Abraham ist in folgender Täuschung begriffen: Die Einförmigkeit dieser Welt kann er nicht ertragen. Nun ist aber die Welt bekanntlich ungemein mannigfaltig, was jederzeit nachzuprüfen ist, indem man eine Handvoll Welt nimmt und näher ansieht.
>
> *Abraham falls victim to the following illusion: he cannot stand the uniformity of this world. Now the world is known, however, to be uncommonly various, which can be verified at any time by taking a handful of the world and looking at it closely.*
>
> — Franz Kafka, "Abraham", transl. Clement Greenberg

In this dissertation, I have investigated how adult speakers and statistical models generalize the famously complex German plural system, with the goal of distinguishing between three proposed criteria for morphological regularity: rule generation, type frequency, and predictability.

I conclude that speaker generalization behavior is best characterized as *probability-matching to a phonologically-conditioned lexical distribution*. This outcome is partly compatible with both the type frequency and predictability accounts of morphological regularity, though not wholly compatible with either; speakers are sensitive to output type frequency and other cues, but largely do not use the predictive cue of grammatical gender, which in principle could substantially reduce uncertainty in plural generalization. This outcome does, however, appear wholly incompatible with any rule generation account which privileges input constraints over output distributions.

In terms of computational modeling, a static phonologically-conditioned lexical baseline most closely approximates speaker productions on average. Beyond this, Bayesian models trained without grammatical gender appear to best capture speakers' probability-matching behavior. Most statistical models rely too heavily on grammatical gender, which only mildly influences speaker generalization.

These findings have broader implications for cognitive models of natural language. The divergence between speaker and model behavior in grammatical gender represents a seemingly rare case in computational linguistics, one where additional data (here, noun gender) impedes rather than facilitates human-like performance. The neural Encoder-Decoder model, along with other learners reviewed in this dissertation, relies upon the strong statistical connection between grammatical gender and plural class — a connection which has informed decades of formal linguistic analysis. In so doing, these models arguably demonstrate "superhuman" plural generalization performance by using a variable to which speakers are insensitive. Many recent capacity breakthroughs have been driven by dramatically increasing the scale of neural network models of natural language; given this trend, I anticipate that future research will reveal many more such divergences, in more complex linguistic domains. Scientists aiming for cognitive models of language use and processing will need to either develop alternative models, or acknowledge and manage the limitations of these statistical learners.

8.1 Contributions

To summarize, the main contributions of this thesis are:

1. A novel conceptual framework of morphological regularity building on Herce (2019), which characterizes key theoretical contrasts between the rule generation, type frequency, and predictability accounts of inflection generalization (§2.1.1).

2. A comprehensive review of theoretical, behavioral, and computational approaches to morphological generalization (Ch. 2), particularly in the domain of German plural inflection (Ch. 3).

3. A series of behavioral experiments using stimuli developed by Marcus et al. (1995) (Chs. 4, 5, 6). The resulting data is published to support reproducibility, and has already been

used by other researchers (Haley, 2020; Beser, 2021; Belth et al., 2021; Dankers et al., 2021; Heitmeier et al., 2021; Rosen, 2022).

4. A critical reanalysis of the data presented by Marcus et al., concluding that their key finding is driven by a stimulus confound, and therefore readily captured by statistical learners (Ch. 4).
5. A series of modeling experiments with different types of statistical learners (Chs. 4, 5, 7) to evaluate whether computational models trained on the German noun lexicon can learn speaker-like plural generalization.
6. A detailed analysis of the cumulative behavioral and computational evidence (Ch. 7), leading to two empirical conclusions:
    (a) Speakers probability-match to a lexical distribution which is primarily conditioned on phonology, and only weakly on grammatical gender.
    (b) Computational models typically do not match the variability of speaker productions, and are outperformed on average by a static phonologically-conditioned lexical baseline.
7. A theoretical conclusion that morphological generalization in this domain is characterized by output distributions and selected input constraints, integrating elements of the type frequency and predictability accounts of morphological regularity.

8.2 Limitations and Future Directions

8.2.1 Behavioral

Stimuli

The four behavioral experiments in this dissertation all used the same small set of 24 stimulus items developed by Marcus et al. (1995). This decision had several advantages. It permitted direct comparison to other experiments which used the same stimuli, both behavioral (Marcus et al., 1995; Zaretsky and Lange, 2016) and computational (Goebel and Indefrey, 2000; Hahn and Nakisa, 2000). It also enabled direct comparison of experimental manipulations within this dissertation, especially the effect of grammatical gender on speakers' plural class productions across different study designs. Limiting the number of stimuli kept experiments short — typically under 10 minutes — preventing dropout and other attention or engagement issues among participants. Finally, having repeated measures for each stimulus item allowed me to quantify the variability in speaker productions across experiments, and conclude with some confidence that speakers consistently generalize particular distributions to these items. These repeated measures lend credibility to the final analysis presented in Chapter 7.

Relying exclusively on these stimuli, however, also significantly limits the conclusions which can be drawn from this thesis. These 24 novel words represent a very small subset of possible German nouns, and deliberately avoid certain phonological and morphological cues with highly predictable plural class associations in the lexicon (c.f. §3.1); it is quite possible that a broader range of stimuli would reveal a broader range of speaker generalization behavior (c.f. Köpcke, 1988; Spreng, 2004). Moreover, while these stimuli are intended to range from phonologically typical (Rhymes) to atypical (Non-Rhymes), there are issues in their design. The most critical issue from a scientific perspective is the *-s-k* confound discussed in Chapter 4, which undermines a central claim made by Marcus et al. (1995); however, other authors have also highlighted phonological and orthographic irregularities (e.g. in responses to Clahsen, 1999*b*), and there are likely further confounds to be discovered. Modern computational tools, such as Wuggy (Keuleers and Brysbaert, 2010) and successors, support the development of a wider range of stimuli with more fine-grained control over confounding factors. Future work in this area could use a broader set of carefully controlled stimuli to better characterize speaker generalization, including the effects of phonological typicality and other phonological and morphological predictive cues.

Cultural transmission

Another limitation of the behavioral experiments reported here is that they focus on how *individual* speakers generalize plural classes on initial exposure to a novel noun. The findings presented in Chapters 5 and 6 show that grammatical gender has a most a weak effect on speakers' initial generalizations. If this is the case, how has gender come to have such a strong statistical association with plural class in the lexicon (§3.1)?

One hypothesis is that gender conditioning could be facilitated by *cultural transmission* across speakers. As discussed in §2.2.2, during interaction or iterated learning across speakers, cultural transmission can exert specific communicative pressures which are absent from spontaneous generalization in isolation; moreover, some experimental evidence suggests that cultural transmission may facilitate conditional regularization in particular, although the findings in this area are inconclusive. Future behavioral work in this area could test whether cultural transmission increases gender conditioning in German plural generalization through experimental designs using either iterated learning across participants (e.g. Smith and Wonnacott, 2010; Smith et al., 2017) or dyadic interaction between participants (e.g. Rácz et al., 2020; Fehér et al., n.d.).

8.2.2 Computational: Toward modeling speaker-like generalization

An open question remains: what kind of computational model *could* learn speaker-like plural generalization from the German noun lexicon? Such a model would presumably require the right set of inductive biases to condition primarily on phonological shape, and only secondarily on grammatical gender. This is challenging for any statistical model, as grammatical gender is a highly predictive cue to plural class.

Locality biases One might imagine that a *locality* or *recency bias* could address this issue: the final syllable of a noun is directly adjacent to its plural suffix, while grammatical gender is expressed at the very beginning of a word. This is not so far away for monosyllabic words like the stimuli considered here, but the notorious compounding tendencies of German nouns can make the grammatical gender to plural class connection a relatively long-range dependency for much of the lexicon. Information locality has been shown to affect various levels of linguistic structure; for example, Hahn et al. (2022) demonstrate that a locality bias can explain morphological affix ordering across a typologically diverse set of languages. It is out of scope for this dissertation to establish whether locality effects account for the relatively stronger influence of word-final phonology (compared to grammatical gender) on plural class generalization; however, we can consider whether explicitly implementing a recency bias could better model this effect. Recurrent Neural Network (RNN) models are structurally biased toward information which has more recently entered their hidden state, so one might imagine that they could capture this preference; the gender focus shown by the Encoder-Decoder models in Chapters 4 and 5 might be attributable to the architectural choice of bi-directional encoders.

Other studies, however, indicate that unidirectional RNNs remain susceptible to the predictive power of grammatical gender. As discussed in §3.4, Goebel and Indefrey (2000) found that simple RNNs with short-term memory effectively learned a gender-conditioned default system for German plural inflection. More recently, Dankers et al. (2021) report a similarly strong bias favoring grammatical gender in the behavior of a unidirectional LSTM (long short-term memory network, i.e. an RNN with a memory cell; Hochreiter and Schmidhuber, 1997).

Modeling variability The issue of neural models' overconfident predictions, highlighted in Chapters 4 and 7, could be addressed by a range of modeling approaches — for instance, conformal prediction intervals (e.g. Angelopoulos and Bates, 2021; Hechtlinger et al., 2019), variational inference and other Bayesian approaches to neural sequence modeling (e.g. Graves, 2011; Chien, 2019), or exploration of various parameter regularization techniques such as dropout. These techniques may better capture the variability in speaker responses, although it is less clear whether they would address the other cognitive modeling challenge, i.e. learning the correct relative factor weightings for grammatical gender and phonological form.

Inflection in context Yet another possibility is that the task formulation used in this dissertation is fundamentally flawed; instead of modeling a mapping from singular to plural forms, we should situate the task of morphological inflection within more realistic contexts — for example, as part of a sentence-level language modeling objective, or within multimodal contexts. This argument has been advanced by Ramscar (2021) and discussed in §5.4. Some recent work has focused on modeling morphological inflection within sentence contexts (e.g. McCarthy et al., 2019; Goldman and Tsarfaty, 2022), and this could be a promising direction for future research. The challenge for such models would be defining an adequate experimental comparison to the behavioral wug test as structured here, in which inflection is presented as a context-free mapping task.

And what about Transformers?

In the course of writing this dissertation, the neural Transformer architecture (Vaswani et al., 2017) has emerged as the top-performing model across most domains of natural language processing (NLP). Despite known limitations in certain tasks (e.g. compositional generalization;

Kim and Linzen, 2020; Soulos et al., 2023), Transformers have driven substantial progress on many NLP benchmarks, including morphological generalization (e.g. Pimentel et al., 2021). In particular, Large Language Models (LLMs) — Transformers trained on vast amounts of linguistic data with a language modeling objective, and typically fine-tuned with additional objectives such as instruction-following — have been claimed to show near human-level performance in a range of domains (Bubeck et al., 2023). While there is no particular reason to believe that Transformer models would have the correct inductive biases to overcome the general statistical challenges raised above, their striking success merits at least brief discussion.

Initial investigations by other researchers suggest that Transformer models are, in fact, susceptible to the same biases seen in RNNs in the domain of German plural inflection. Beser (2021) conducts the most comparable investigation in this regard, training multiple Transformer and RNN model instances on the CELEX German noun lexicon. He finds that the Transformer is equally sensitive to grammatical gender, and does not appear to better approximate the human production data collected by McCurdy, Goldwater and Lopez (2020, Ch. 4). Haley (2020) uses the same novel nouns to probe number agreement in German and multilingual versions of BERT (Devlin et al., 2018), a Transformer Encoder-Decoder architecture pretrained using a Masked Language Modeling objective — a precursor to modern LLMs. He finds that the models perform perfectly for real nouns but drop below 80% accuracy for novel nouns, suggesting difficulty in processing novel plural forms.

One particular issue that arises in modern LLM evaluation is the difficulty of assessing out-of-distribution generalization; with models trained on internet-scale data, it is nearly impossible to confirm that the model hasn't been exposed to a particular test item during training (Hupkes et al., 2022). We can safely assume that the vast majority of human speakers have not encountered the novel stimuli developed by Marcus et al. (1995), but we cannot make that same assumption for LLMs, which are often exposed to scientific literature and related writing in their training data. To illustrate this point, I offer several selected responses from the Bing chatbot, a commercial application powered by the GPT-4 LLM which has shown impressive abilities (Bubeck et al., 2023). German-language prompting reveals that the Marcus et al. stimuli are readily available on the internet and therefore likely appear in LLM training corpora.

> Ich kann Ihnen leider nicht sagen, wie viele Plaupfe es auf der Welt gibt, denn Plaupf ist kein echtes deutsches Wort. Es ist ein erfundenes Wort, das von einigen

> Forschern verwendet wurde, um zu testen, wie deutsche Sprecher neue Wörter beugen. Es scheint auch kein Wort in einer anderen Sprache zu sein, denn meine Suche nach Plaupf hat nur Ergebnisse über Sport und die Bundesliga ergeben. Vielleicht haben Sie sich das Wort ausgedacht oder es irgendwo gelesen?
>
> Source: Conversation with Bing, 01/05/2023
>
> *Unfortunately, I cannot tell you how many Plaupfe there are in the world, because that is not a real German word. It was made up by some researchers to test how German speakers inflect new words. It also seems not to be a word in any other language, as my search for Plaupf has only found results about sports and the Bundesliga. Maybe you thought up this word, or read it somewhere?*
>
> Es gibt verschiedene Forscher, die solche erfundenen Wörter studieren, um zu verstehen, wie Menschen neue Wörter lernen und verwenden. Zum Beispiel haben Kate McCurdy, Sharon Goldwater und Adam Lopez eine Studie durchgeführt, in der sie deutsche Sprecher gebeten haben, die Pluralformen von erfundenen Wörtern wie Plaupf zu produzieren. Sie haben auch ein neuronales Netzwerk trainiert, um das gleiche zu tun, und die Ergebnisse verglichen. Sie haben herausgefunden, dass das neuronale Netzwerk sich anders verhält als die Sprecher und nicht die seltenen Muster anwendet, die die Sprecher verwenden. Das heißt, dass das neuronale Netzwerk nicht vollständig das menschliche Sprachverhalten nachahmt.
>
> Source: Conversation with Bing, 01/05/2023
>
> *There are various researchers who study such invented words to understand how people learn and use new words. For example, Kate McCurdy, Sharon Goldwater and Adam Lopez conducted a study in which they asked German speakers to produce the plural forms of invented words like Plaupf. They also trained a neural network to do the same, and compared the results. They discovered that the neural network behaved differently, and didn't use the rare patterns applied by speakers. This means the neural network did not fully imitate human speaker behavior.*

Based on some further prompting with other novel nouns, the Bing chatbot remains inclined to assign plural classes on the basis of grammatical gender and final character of the singular noun, even citing specific (though invented) analogous examples. It does, however, generate word sequences which appear to acknowledge the fundamentally variable nature of German plural generalization.

> Das ist auch eine mögliche Pluralform. Fnahfe klingt wie ein Muster, das für einige feminine Wörter gilt, die auf -f enden, wie zum Beispiel Scharf - Scharfe oder Scherf - Scherfe. Vielleicht empfinden Sie Fnahf als feminin und nicht als neutral. Oder vielleicht haben Sie eine andere Assoziation zu dem Wort. Wie ich schon gesagt habe, gibt es keine feste Regel für erfundene Wörter, sondern nur Tendenzen und Präferenzen. Das macht die deutsche Sprache so spannend und vielfältig.
>
> Source: Conversation with Bing, 01/05/2023
>
> *That is also a possible plural form. Fnahfe sounds like a pattern that applies for several feminine words that end in -f, for example Scharf - Scharfe [N.B. this is incorrect: the adjective scharf inflects as scharfe to agree with feminine nouns, but is not a feminine noun itself -K] or Scherf - Scherfe [N.B. Scherf is not a word in German -K] . Perhaps you feel that Fnahf is feminine and not neuter, or perhaps you have some other associations with this word. As I said before, there are no definite rules for made-up words, only tendencies and preferences. This makes the German language so diverse and exciting.*

Diverse and exciting indeed.

8.3 Finis

The past five years have seen significant developments in natural language processing technology. The historical record shows both social upheaval and adaptation in response to previous disruptive cultural technologies, such as the internet, the printing press, and the written word. It now appears that we may being entering a comparable yet wholly new era: only in recent years have we faced the prospect of human-like natural language interaction with non-human technological entities, such as the chatbot quoted above. Along with all the other questions this raises, there is an opportunity to revisit core scientific questions. What, in fact, specifically characterizes human natural language use?

In this dissertation, I have chosen to focus on a highly restricted problem domain from a three-decade-old debate, in hopes of shedding a small bit of light onto this question. The findings reported here point toward a minor, relatively inconsequential divergence between certain artificially constrained measures of human linguistic behavior, and comparably constrained measures of computational linguistic behavior. Nonetheless, my hunch is that powerful statistical models of language will point us toward many such divergences in the future, and these findings will help us clarify the nature of human language processing. We will learn to conceptually decouple linguistic and cognitive phenomena that until now only came bundled together, in human form.

References

Ackerman, F., Blevins, J. P. and Malouf, R. (2009), 'Parts and wholes: Implicative patterns in inflectional paradigms', *Analogy in grammar: Form and acquisition* pp. 54–82. Publisher: Oxford University Press Oxford.

Ackerman, F. and Malouf, R. (2013), 'Morphological Organization: The Low Conditional Entropy Conjecture', *Language* **89**(3), 429–464. Publisher: Linguistic Society of America.
URL: *https://muse.jhu.edu/article/521667*

Albright, A. and Hayes, B. (2003), 'Rules vs. analogy in English past tenses: a computational/experimental study', *Cognition* **90**(2), 119–161.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S001002770300146X*

Ambridge, B. (2020), 'Against stored abstractions: A radical exemplar model of language acquisition', *First Language* **40**(5-6), 509–559. Publisher: SAGE Publications Ltd.
URL: *https://doi.org/10.1177/0142723719869731*

Anderson, S. R. (1982), 'Where's Morphology?', *Linguistic Inquiry* **13**(4), 571–612. Publisher: MIT Press.
URL: *https://www.jstor.org/stable/4178297*

Angelopoulos, A. N. and Bates, S. (2021), 'A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification', *arXiv:2107.07511 [cs, math, stat]* . arXiv: 2107.07511.
URL: *http://arxiv.org/abs/2107.07511*

Ashby, F. G. and Rosedahl, L. (2017), 'A neural interpretation of exemplar theory.', *Psychological Review* **124**(4), 472. Publisher: US: American Psychological Association.
URL: *https://psycnet.apa.org/fulltext/2017-15496-001.pdf*

Augst, G. (1979), 'Neuere Forschungen zur Substantivflexion', *Zeitschrift für Germanistische Linguistik* **7**, 220. Place: Berlin Publisher: Walter de Gruyter.
URL: *https://www.proquest.com/scholarly-journals/neuere-forschungen-zur-substantivflexion/docview/1303724383/se-2?accountid=10673*

Austin, A. C. (2010), When children learn more than what they are taught: Regularization in child and adult learners, Ph.D., University of Rochester, United States – New York. ISBN: 9781124303444.
URL: *https://www.proquest.com/docview/807620951/abstract/6FDDCC59FD9F442FPQ/1*

Baayen, R. H. (2011), 'Corpus linguistics and naive discriminative learning', *Revista Brasileira de Linguística Aplicada* **11**, 295–328. Publisher: Faculdade de Letras - Universidade Federal de Minas Gerais.
URL: *http://www.scielo.br/j/rbla/a/fjpYkMycJdW8xypVYzbzWDq/?format=htmllang=en*

Baayen, R. H. and Hendrix, P. (2017), 'Two-layer networks, non-linear separation, and human learning', *From Semantics to Dialectometry. Festschrift in honor of John Nerbonne. Tributes* **32**, 13–22.

Baayen, R. H., Piepenbrock, R. and Gulikers, L. (1995), 'The CELEX2 lexical database', *Linguistic Data Consortium* .

Baroni, M. (2021), 'On the proper role of linguistically-oriented deep net analysis in linguistic theorizing', *arXiv:2106.08694 [cs]* . arXiv: 2106.08694.
URL: *http://arxiv.org/abs/2106.08694*

Baroni, M. (2022), On the Proper Role of Linguistically Oriented Deep Net Analysis in Linguistic Theorising, *in* 'Algebraic Structures in Natural Language', CRC Press, pp. 1–16.

Bartke, S., Marcus, G. and Clahsen, H. (1995), Acquiring German noun plurals, *in* 'Proceedings of the 19th Annual Boston University Conference on Language Development', Cascadilla Press.

Bartsch, R. and Vennemann, T. (1983), 'Grundzüge der Sprachtheorie: Eine Linguistische Einfuerung', *Principles of language theory: a linguistic introduction.) Tübingen: Max Niemeyer* .

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015), 'Fitting Linear Mixed-Effects Models Using lme4', *Journal of Statistical Software* **67**(1), 1–48.

Bauer, L. (2001), *Morphological Productivity*, Cambridge Studies in Linguistics, Cambridge University Press, Cambridge.
URL: *https://www.cambridge.org/core/books/morphological-productivity/DE18903CDF643BE23740D3ABF341831F*

Becker, M., Ketrez, N. and Nevins, A. (2011), 'THE SURFEIT OF THE STIMULUS: ANALYTIC BIASES FILTER LEXICAL STATISTICS IN TURKISH LARYNGEAL ALTERNATIONS', *Language* **87**(1), 84–125. Publisher: Linguistic Society of America.
URL: *https://www.jstor.org/stable/23011541*

Belth, C. A., Payne, S. R. B., Beser, D., Kodner, J. and Yang, C. (2021), The Greedy and Recursive Search for Morphological Productivity, *in* 'Proceedings of the Annual Meeting of the Cognitive Science Society', Vol. 43.
URL: *https://escholarship.org/uc/item/1md2p6j5*

Bentz, C. and Ferrer Cancho, R. (2016), Zipf's law of abbreviation as a language universal, University of Tübingen, pp. 1–4. Accepted: 2020-02-27T15:47:46Z.
URL: *https://upcommons.upc.edu/handle/2117/178845*

Berko, J. (1958), 'The Child's Learning of English Morphology', *WORD* **14**(2-3), 150–177.
URL: *http://www.tandfonline.com/doi/full/10.1080/00437956.1958.11659661*

Beser, D. (2021), Falling Through the Gaps: Neural Architectures as Models of Morphological Rule Learning, *in* 'Proceedings of the Annual Meeting of the Cognitive Science Society', Vol. 43.
URL: *https://escholarship.org/uc/item/1dp545wm*

Bittner, A. and Köpcke, K.-M. (2016), *Regularität und Irregularität in Phonologie und Morphologie : Diachron, Kontrastiv, Typologisch*, Walter de Gruyter GmbH, Berlin/Boston, UNITED STATES.
URL: *http://ebookcentral.proquest.com/lib/ed/detail.action?docID=4768975*

Bittner, D. (1994), 'Die Bedeutung der Genusklassifikation für die Organisation der deutschen Substantivflexion', *Funktionale Untersuchungen zur deutschen Nominal-und Verbalmorphologie* pp. 65–80.

Bittner, D. (1999), Gender classification and the inflectional system of German nouns, *in* 'Gender in Grammar and Cognition: I: Approaches to Gender. II: Manifestations of Gender', number 124 *in* 'Trends in linguistics', De Gruyter Mouton, Berlin ; New York, pp. 1–24.
URL: *https://doi.org/10.1515/9783110802603*

Blevins, J. P., Milin, P. and Ramscar, M. (2017), The Zipfian Paradigm Cell Filling Problem, *in* F. Kiefer, J. Blevins and H. Bartos, eds, 'Perspectives on Morphological Organization', BRILL, pp. 139–158.
URL: *https://brill.com/view/book/edcoll/9789004342934/B9789004342934*$_0 08.xml$

Bloomfield, L. (1933), *Language*, Motilal Banarsidass Publ. reissued edition 1994.

Blything, R. P., Ambridge, B. and Lieven, E. V. (2018), 'Children's Acquisition of the English Past-Tense: Evidence for a Single-Route Account From Novel Verb Production Data', *Cognitive Science* **42**, 621–639.
URL: *http://doi.wiley.com/10.1111/cogs.12581*

Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017), 'Enriching Word Vectors with Subword Information', *Transactions of the Association for Computational Linguistics* **5**, 135–146. Place: Cambridge, MA Publisher: MIT Press.
URL: *https://aclanthology.org/Q17-1010*

Booij, G. (2016), Construction Morphology, *in* A. Hippisley and G. Stump, eds, 'The Cambridge Handbook of Morphology', Cambridge University Press, Cambridge, pp. 424–448.
URL: *https://www.cambridge.org/core/product/identifier/9781139814720*

Brown, D. (2016), Defaults and Overrides in Morphological Description, *in* A. Hippisley and G. Stump, eds, 'The Cambridge Handbook of Morphology', 1 edn, Cambridge University Press, pp. 272–296.
URL: *https://www.cambridge.org/core/product/identifier/9781139814720*

Brown, H., Smith, K., Samara, A. and Wonnacott, E. (2022), 'Semantic cues in language learning: an artificial language study with adult and child learners', *Language, Cognition and Neuroscience* **37**(4), 509–531. Publisher: Routledge _eprint: https://doi.org/10.1080/23273798.2021.1995612.
URL: *https://doi.org/10.1080/23273798.2021.1995612*

Brown, R. (1973), *A first language: The early stages*, Harvard University Press.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T. and Zhang, Y. (2023), 'Sparks of Artificial General Intelligence: Early experiments with GPT-4'. arXiv:2303.12712 [cs].
URL: *http://arxiv.org/abs/2303.12712*

Bybee, J. (1995), 'Regular morphology and the lexicon', *Language and Cognitive Processes* **10**(5), 425–455.
URL: *http://www.tandfonline.com/doi/abs/10.1080/01690969508407111*

Bybee, J. (1996), 'Productivity, regularity and fusion: how language use affects the lexicon', *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4* pp. 247–269.

Bybee, J. (2006*a*), *Frequency of use and the organization of language*, Oxford University Press.

Bybee, J. (2010), *Language, usage and cognition*, Cambridge University Press.

Bybee, J. L. (2006*b*), 'From Usage to Grammar: The Mind's Response to Repetition', *Language* **82**(4), 711–733.
URL: *http://muse.jhu.edu/content/crossref/journals/language/v082/82.4bybee.pdf*

Bybee, J. L. and Slobin, D. I. (1982), 'Rules and Schemas in the Development and Use of the English past Tense', *Language* .

Bürkner, P.-C. (2017), '**brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*', *Journal of Statistical Software* **80**(1).
URL: *http://www.jstatsoft.org/v80/i01/*

Bürkner, P.-C. (2020), 'Bayesian Item Response Modeling in R with brms and Stan', *arXiv:1905.09501 [stat]* . arXiv: 1905.09501.
URL: *http://arxiv.org/abs/1905.09501*

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017), 'Stan: A probabilistic programming language', *Journal of statistical software* **76**(1). Publisher: Columbia Univ., New York, NY (United States); Harvard Univ., Cambridge, MA (United States).

Chambers, J. K. and Schilling, N. (2018), *The handbook of language variation and change*, John Wiley & Sons.

Chandler, S. (2010), 'The English past tense: Analogy redux', **21**(3), 371–417. Publisher: De Gruyter Mouton Section: Cognitive Linguistics.
URL: *https://www.degruyter.com/document/doi/10.1515/COGL.2010.014/html*

Chandler, S. (2017), 'The analogical modeling of linguistic categories', *Language and Cognition* **9**(1), 52–87. Publisher: Cambridge University Press.
URL: *https://www.cambridge.org/core/journals/language-and-cognition/article/analogical-modeling-of-linguistic-categories/83168A8F78C16AB07ADD300B2F459D57*

Chien, J.-T. (2019), Deep Bayesian Natural Language Processing, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts', Association for Computational Linguistics, Florence, Italy, pp. 25–30.
URL: *https://aclanthology.org/P19-4006*

Chomsky, N. and Halle, M. (1968), 'The Sound Pattern of English', *undefined* .
URL: */paper/The-Sound-Pattern-of-English-Chomsky-Halle/2ff60c8f2df9a57b5dfba38e9be6d6401d88a1e9*

Clahsen, H. (1999*a*), 'The dual nature of the language faculty', *Behavioral and Brain Sciences* **22**(6), 1046–1055.
URL: *http://www.journals.cambridge.org/abstract*$_S0140525X99562225$

Clahsen, H. (1999*b*), 'Lexical entries and rules of language: A multidisciplinary study of German inflection', *Behavioral and Brain Sciences* **22**(6), 991–1013.
URL: *https://www.cambridge.org/core/product/identifier/S0140525X99002228/type/journal*$_article$

Clahsen, H. (2016), Experimental Studies of Morphology and Morphological Processing, *in* 'The Cambridge Handbook of Morphology', 1 edn, Cambridge University Press, pp. 792–819.
URL: *https://www.cambridge.org/core/product/identifier/9781139814720*

Corkery, M., Matusevych, Y. and Goldwater, S. (2019), Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection, *in* 'Proceedings of the 57th Conference of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 3868–3877.
URL: *https://www.aclweb.org/anthology/P19-1376*

Cotterell, R., Kirov, C., Hulden, M. and Eisner, J. (2018), 'On the complexity and typology of inflectional morphological systems', *Transactions of the Association for Computational Linguistics* **7**, 327–342.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., McCarthy, A. D., Kann, K., Mielke, S. J., Nicolai, G., Silfverberg, M., Yarowsky, D., Eisner, J. and Hulden, M. (2018), The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection, *in* 'Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection', Association for Computational Linguistics, Brussels, pp. 1–27.
URL: *https://www.aclweb.org/anthology/K18-3001*

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Kübler, S., Yarowsky, D., Eisner, J. and Hulden, M. (2017), CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages, *in* 'Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection', Association for Computational Linguistics, Vancouver, pp. 1–30.
URL: *http://aclweb.org/anthology/K17-2001*

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J. and Hulden, M. (2016), The SIGMORPHON 2016 Shared Task—Morphological Reinflection, *in* 'Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology', Association for Computational Linguistics, Berlin, Germany, pp. 10–22.
URL: *http://aclweb.org/anthology/W16-2002*

Culbertson, J., Gagliardi, A. and Smith, K. (2017), 'Competition between phonological and semantic cues in noun class learning', *Journal of Memory and Language* **92**, 343–358.
URL: *http://www.sciencedirect.com/science/article/pii/S0749596X16300687*

Culbertson, J. and Kirby, S. (2016), 'Simplicity and Specificity in Language: Domain-General Biases Have Domain-Specific Effects', *Frontiers in Psychology* **6**.
URL: *https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01964*

Culbertson, J. and Schuler, K. (2019), 'Artificial Language Learning in Children', *Annual Review of Linguistics* **5**(1), 353–373.
URL: *https://www.annualreviews.org/doi/10.1146/annurev-linguistics-011718-012329*

Daelemans, W. (2002), 'A comparison of analogical modeling to memory-based language processing', *Skousen et al* pp. 157–179.

Dankers, V., Langedijk, A., McCurdy, K., Williams, A. and Hupkes, D. (2021), Generalising to German Plural Noun Classes, from the Perspective of a Recurrent Neural Network, *in* 'Proceedings of the 25th Conference on Computational Natural Language Learning', Association for Computational Linguistics, Online, pp. 94–108.
URL: *https://aclanthology.org/2021.conll-1.8*

Dasgupta, I., Grant, E. and Griffiths, T. L. (2021), 'Distinguishing rule- and exemplar-based generalization in learning systems', *arXiv:2110.04328 [cs]* . arXiv: 2110.04328.
URL: *http://arxiv.org/abs/2110.04328*

Dawdy-Hesterberg, L. G. (2014), The Structural and Statistical Basis of Morphological Generalization in Arabic, Ph.D., Northwestern University, United States – Illinois. ISBN: 9781321446678.
URL: *https://www.proquest.com/docview/1648654673/abstract/A93BC63BCB554A9FPQ/1*

Deo, A. (2007), 'Derivational morphology in inheritance-based lexica: Insights from Pāṇini', *Lingua* **117**(1), 175–201.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0024384106000167*

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018), 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *arXiv:1810.04805 [cs]* . arXiv:

1810.04805.

URL: *http://arxiv.org/abs/1810.04805*

Dressler, W. U., Mayerthaler, W., Panagl, O. and Wurzel, W. U. (1987), *Leitmotifs in natural morphology*, Benjamins Amsterdam.

Dye, M., Milin, P., Futrell, R. and Ramscar, M. (2017), A Functional Theory of Gender Paradigms, *in* F. Kiefer, J. Blevins and H. Bartos, eds, 'Perspectives on Morphological Organization', Brill, pp. 212–239.

URL: *http://booksandjournals.brillonline.com/content/books/b9789004342934*$_0$11

Ernestus, M. T. C. and Baayen, R. H. (2003), Predicting the Unpredictable: Interpreting Neutralized Segments in Dutch, *in* 'Language', Vol. 79, pp. 5–38. Publisher: Linguistic Society of America.

URL: *https://muse.jhu.edu/pub/24/article/41846*

Ettlinger, M., Morgan-Short, K., Faretta-Stutenberg, M. and Wong, P. C. (2016), 'The Relationship Between Artificial and Second Language Learning', *Cognitive Science* **40**(4), 822–847. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12257.

URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12257*

Feder, M. (1986), 'Maximum entropy as a special case of the minimum description length criterion (Corresp.)', *IEEE Transactions on Information Theory* **32**(6), 847–849. Conference Name: IEEE Transactions on Information Theory.

Fehér, O., Järvinen, H., Wonnacott, E. and Smith, K. (n.d.), 'The effect of interaction on the evolution of conditioned variation'.

Fehér, O., Wonnacott, E. and Smith, K. (2016), 'Structural priming in artificial languages and the regularisation of unpredictable variation', *Journal of Memory and Language* **91**, 158–180.

URL: *https://www.sciencedirect.com/science/article/pii/S0749596X1630047X*

Feldman, N. (2005), Learning and overgeneralization patterns in a connectionist model of the German plural, Master's thesis, University of Vienna.

Ferdinand, V., Kirby, S. and Smith, K. (2019), 'The cognitive roots of regularization in language', *Cognition* **184**, 53–68.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0010027718303135*

Ferman, S. and Karni, A. (2010), 'No Childhood Advantage in the Acquisition of Skill in Using an Artificial Language Rule', *PLOS ONE* **5**(10), e13648. Publisher: Public Library of Science.
URL: *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0013648*

Frigo, L. and McDonald, J. L. (1998), 'Properties of Phonological Markers That Affect the Acquisition of Gender-Like Subclasses', *Journal of Memory and Language* **39**(2), 218–245.
URL: *http://www.sciencedirect.com/science/article/pii/S0749596X98925694*

Gaeta, L. (2008), 'Die deutsche Pluralbildung zwischen deskriptiver Angemessenheit und Sprachtheorie*', *Zeitschrift für Germanistische Linguistik* **36**(1), 74–108. Place: Berlin Publisher: Walter de Gruyter GmbH.
URL: *https://www.proquest.com/scholarly-journals/die-deutsche-pluralbildung-zwischen-deskriptiver/docview/204448880/se-2?accountid=10673*

Gagliardi, A., Feldman, N. H. and Lidz, J. (2017), 'Modeling Statistical Insensitivity: Sources of Suboptimal Behavior', *Cognitive Science* **41**(1), 188–217.
URL: *https://onlinelibrary.wiley.com/doi/10.1111/cogs.12373*

Gagliardi, A. and Lidz, J. (2014), 'Statistical Insensitivity in the Acquisition of Tsez Noun Classes', *Language* **90**(1), 58–89. Publisher: Linguistic Society of America.
URL: *https://www.jstor.org/stable/24672020*

Gardner, R. A. (1957), 'Probability-Learning with Two and Three Choices', *The American Journal of Psychology* **70**(2), 174–185. Publisher: University of Illinois Press.
URL: *https://www.jstor.org/stable/1419319*

Gawlitzek-Maiwald, I. (1994), How do children cope with variation in the input? The case of German plurals and compounding, *in* 'How tolerant is universal grammar? Essays on language learnability and language variation', number 309 *in* 'Linguistische Arbeiten', Max Niemeyer, Tübingen, pp. 225–266.

Gelhaus, H. (1998), ‘Die Wortarten: Die Flektierbaren und die Unflektierbaren’, *Duden. Grammatik der deutschen Gegenwartssprache. Duden* **4**, 6.

Genz, A. and Bretz, F. (2009), *Computation of Multivariate Normal and t Probabilities*, Lecture Notes in Statistics, Springer-Verlag, Heidelberg.

Gisborne, N. (2017), Defaulting to the new Romance synthetic future, *in* N. Gisborne and A. Hippisley, eds, ‘Defaults in Morphological Theory’, Oxford University Press, p. 0.
URL: *https://doi.org/10.1093/oso/9780198712329.003.0007*

Gisborne, N. and Hippisley, A. (2017), Defaults in linguistics, *in* N. Gisborne and A. Hippisley, eds, ‘Defaults in Morphological Theory’, Oxford University Press, p. 0.
URL: *https://doi.org/10.1093/oso/9780198712329.003.0001*

Goebel, R. and Indefrey, P. (2000), ‘A recurrent network with short-term memory capacity learning the German-s plural’, *Models of language acquisition: Inductive and deductive approaches* pp. 177–200.

Goldberg, A. E. (2006), *Constructions at work : the nature of generalization in language /*, Oxford University Press,, Oxford ; New York :.

Goldman, O., Guriel, D. and Tsarfaty, R. (2021), ‘(Un)solving Morphological Inflection: Lemma Overlap Artificially Inflates Models’ Performance’, *arXiv:2108.05682 [cs]* . arXiv: 2108.05682.
URL: *http://arxiv.org/abs/2108.05682*

Goldman, O. and Tsarfaty, R. (2022), ‘Morphology Without Borders: Clause-Level Morphology’. arXiv:2202.12832 [cs].
URL: *http://arxiv.org/abs/2202.12832*

Goldwater, S. and Johnson, M. (2003), Learning OT Constraint Rankings Using a Maximum Entropy Model, *in* ‘Proceedings of the Stockholm Workshop on Variation within Optimality Theory’, pp. 111–120.

Graves, A. (2011), Practical Variational Inference for Neural Networks, *in* ‘Advances in Neural Information Processing Systems’, Vol. 24, Curran Associates, Inc.

URL: *https://proceedings.neurips.cc/paper*$_{f}$*iles/paper/2011/hash/7eb3c8be3d411e8ebfab08eba5f49632—Abstract.html*

Grünwald, P. D. (2007), *The Minimum Description Length Principle*, MIT Press. Google-Books-ID: mbU6T7oUrBgC.

Hahn, M., Mathew, R. and Degen, J. (2022), 'Morpheme Ordering Across Languages Reflects Optimization for Processing Efficiency', *Open Mind* **5**, 208–232.
URL: *https://doi.org/10.1162/opmi*$_{a0}$0051

Hahn, U. (1999), 'The dual-route account of German: Where it is not a schema theory, it is probably wrong', *Behavioral and Brain Sciences* **22**(6), 1024–1025.

Hahn, U. and Chater, N. (1998), 'Similarity and rules: distinct? exhaustive? empirically distinguishable?', *Cognition* **65**(2-3), 197–230.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0010027797000449*

Hahn, U. and Nakisa, R. C. (2000), 'German Inflection: Single Route or Dual Route?', *Cognitive Psychology* **41**(4), 313–360.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0010028500907370*

Haley, C. (2020), This is a BERT. Now there are several of them. Can they generalize to novel words?, *in* 'Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP', Association for Computational Linguistics, Online, pp. 333–341.
URL: *https://www.aclweb.org/anthology/2020.blackboxnlp-1.31*

Haspelmath, M. (2021), 'Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability', *Journal of Linguistics* **57**(3), 605–633. Publisher: Cambridge University Press.
URL: *https://www.cambridge.org/core/journals/journal-of-linguistics/article/explaining-grammatical-coding-asymmetries-formfrequency-correspondences-and-predictability/420965EC1CEA49527CCE7276B33A14D0*

Hayes, B., Siptár, P., Zuraw, K. and Londe, Z. (2009), 'Natural and Unnatural Constraints in Hungarian Vowel Harmony', *Language* **85**(4), 822–863. Publisher: Linguistic Society of

America.
URL: *https://www.jstor.org/stable/40492955*

Heafield, K. (2011), KenLM: Faster and Smaller Language Model Queries, *in* 'Proceedings of the Sixth Workshop on Statistical Machine Translation', Association for Computational Linguistics, Edinburgh, Scotland, pp. 187–197.
URL: *https://aclanthology.org/W11-2123*

Hechtlinger, Y., Póczos, B. and Wasserman, L. (2019), 'Cautious Deep Learning'. arXiv:1805.09460 [cs, stat].
URL: *http://arxiv.org/abs/1805.09460*

Heitmeier, M., Chuang, Y.-Y. and Baayen, R. H. (2021), 'Modeling morphology with Linear Discriminative Learning: considerations and design choices', *arXiv:2106.07936 [cs]* . arXiv: 2106.07936.
URL: *http://arxiv.org/abs/2106.07936*

Heitmeier, M. and Frank, S. (2021), Exploring the influence of semantics on the German plural system: a wug study, *in* 'Proceedings of the Annual Meeting of the Cognitive Science Society', Vol. 43.
URL: *https://escholarship.org/uc/item/8jh84047*

Hendricks, A. E., Miller, K. and Jackson, C. N. (2018), 'Regularizing Unpredictable Variation: Evidence from a Natural Language Setting', *Language Learning and Development* **14**(1), 42–60.
URL: *https://www.tandfonline.com/doi/full/10.1080/15475441.2017.1340842*

Herce, B. (2019), 'Deconstructing (ir)regularity', *Studies in Language* **43**(1), 44–91.
URL: *http://www.jbe-platform.com/content/journals/10.1075/sl.17042.her*

Hochreiter, S. and Schmidhuber, J. (1997), 'Long short-term memory', *Neural computation* **9**(8), 1735–1780.

Hudson Kam, C. L. (2015), 'The impact of conditioning variables on the acquisition of variation in adult and child learners', *Language* **91**(4), 906–937. Publisher: Linguistic Society of

America.
URL: *https://muse.jhu.edu/article/604079*

Hudson Kam, C. L. (2019), 'Reconsidering Retrieval Effects on Adult Regularization of Inconsistent Variation in Language', *Language Learning and Development* **15**(4), 317–337.
URL: *https://www.tandfonline.com/doi/full/10.1080/15475441.2019.1634575*

Hudson Kam, C. L. and Newport, E. L. (2005), 'Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change', *Language Learning and Development* **1**(2), 151–195.
URL: *http://www.tandfonline.com/doi/abs/10.1080/15475441.2005.9684215*

Hudson Kam, C. L. and Newport, E. L. (2009), 'Getting it right by getting it wrong: When learners change languages', *Cognitive Psychology* **59**(1), 30–66.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0010028509000048*

Hupkes, D., Giulianelli, M., Dankers, V., Artetxe, M., Elazar, Y., Pimentel, T., Christodoulopoulos, C., Lasri, K., Saphra, N., Sinclair, A., Ulmer, D., Schottmann, F., Batsuren, K., Sun, K., Sinha, K., Khalatbari, L., Ryskina, M., Frieske, R., Cotterell, R. and Jin, Z. (2022), 'State-of-the-art generalisation research in NLP: a taxonomy and review'. arXiv:2210.03050 [cs].
URL: *http://arxiv.org/abs/2210.03050*

Icard, T. (2021), 'Why Be Random?', *Mind* **130**(517), 111–139.
URL: *https://doi.org/10.1093/mind/fzz065*

Indefrey, P. (1999), 'Some problems with the lexical status of nondefault inflection', *Behavioral and Brain Sciences* **22**(6), 1025–1025.

Janda, R. D. (1990), Frequency, markedness and morphological change: on predicting the spread of noun-plural-s in Modern High German and West Germanic, *in* 'Proceedings of the Eastern States Conference on Linguistics (ESCOL)', Vol. 7, ERIC, pp. 136–153.

Johnson, M. (2017), 'Marr's levels and the minimalist program', *Psychonomic Bulletin & Review* **24**(1), 171–174.
URL: *https://doi.org/10.3758/s13423-016-1062-1*

Johnson, M., Griffiths, T. L. and Goldwater, S. (2007), Adaptor Grammars:A Framework for Specifying Compositional Nonparametric Bayesian Models, *in* B. Schölkopf, J. Platt and T. Hofmann, eds, 'Advances in Neural Information Processing Systems 19', The MIT Press.
URL: *https://direct.mit.edu/books/book/3168/chapter/87467/adaptor-grammars-a-framework-for-specifying*

Johnson, T., Culbertson, J., Rabagliati, H. and Smith, K. (2020), Assessing Integrative Complexity as a predictor of morphological learning using neural networks and artificial language learning, preprint, PsyArXiv.
URL: *https://osf.io/yngw9*

Kann, K. and Schütze, H. (2016), Single-Model Encoder-Decoder with Explicit Morphological Representation for Reinflection, *in* 'Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)', Association for Computational Linguistics, Berlin, Germany, pp. 555–560.
URL: *http://aclweb.org/anthology/P16-2090*

Kawahara, S. (2015), 'Comparing a forced-choice wug test and a naturalness rating test: an exploration using rendaku', *Language Sciences* **48**, 42–47.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0388000114001491*

Kemp, C., Perfors, A. and Tenenbaum, J. B. (2007), 'Learning overhypotheses with hierarchical Bayesian models', *Developmental Science* **10**(3), 307–321. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-7687.2007.00585.x.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-7687.2007.00585.x*

Keuleers, E. and Brysbaert, M. (2010), 'Wuggy: A multilingual pseudoword generator', *Behavior Research Methods* **42**(3), 627–633.
URL: *http://link.springer.com/10.3758/BRM.42.3.627*

Keuleers, E., Sandra, D., Daelemans, W., Gillis, S., Durieux, G. and Martens, E. (2007), 'Dutch plural inflection: The exception that proves the analogy', *Cognitive Psychology* **54**(4), 283–318.
URL: *https://www.sciencedirect.com/science/article/pii/S0010028506000582*

Kharitonov, E. and Chaabouni, R. (2020), What they do when in doubt: a study of inductive biases in seq2seq learners.

URL: *https://openreview.net/forum?id=YmA86Zo-P_t*

Kihm, A. (2017), Old French declension: A Word and Paradigm approach and the role of syncretisms and defaults in its rise and fall, *in* N. Gisborne and A. Hippisley, eds, 'Defaults in Morphological Theory', Oxford University Press, p. 0.

URL: *https://doi.org/10.1093/oso/9780198712329.003.0003*

Kim, N. and Linzen, T. (2020), COGS: A Compositional Generalization Challenge Based on Semantic Interpretation, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Online, pp. 9087–9105.

URL: *https://aclanthology.org/2020.emnlp-main.731*

Kiparsky, P. (1973), Elsewhere in phonology., *in* S. Anderson and P. Kiparsky, eds, 'A Festschrift for Morris Halle', New York: Holt, Rinehart and Winston, pp. 93–106.

Kiparsky, P. (1982), Lexical morphology and phonology, *in* I.-S. Yang, ed., 'Linguistics in the Morning Calm', Hanshin, Seoul, pp. 3–9.

Kirby, S. (2001), 'Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity', *IEEE Transactions on Evolutionary Computation* **5**(2), 102–110. Conference Name: IEEE Transactions on Evolutionary Computation.

Kirby, S., Griffiths, T. and Smith, K. (2014), 'Iterated learning and the evolution of language', *Current Opinion in Neurobiology* **28**, 108–114.

URL: *https://www.sciencedirect.com/science/article/pii/S0959438814001421*

Kirov, C. and Cotterell, R. (2018), 'Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker and Prince (1988) and the Past Tense Debate', *Transactions of the Association for Computational Linguistics* **6**, 651–665.

URL: *https://aclanthology.org/Q18-1045/*

Kirov, C., Sylak-Glassman, J., Que, R. and Yarowsky, D. (2016), Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms, *in* 'Proceedings of the Tenth

International Conference on Language Resources and Evaluation (LREC 2016)', European Language Resources Association (ELRA), Portorož, Slovenia, pp. 3121–3126.
URL: *https://www.aclweb.org/anthology/L16-1498*

Kodner, J. and Khalifa, S. (2022), SIGMORPHON–UniMorph 2022 Shared Task 0: Modeling Inflection in Language Acquisition, *in* 'Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology', Association for Computational Linguistics, Seattle, Washington, pp. 157–175.
URL: *https://aclanthology.org/2022.sigmorphon-1.18*

Köpcke, K.-M. (1988), 'Schemas in German plural formation', *Lingua* **74**(4), 303–335.
URL: *http://linkinghub.elsevier.com/retrieve/pii/0024384188900642*

Köpcke, K.-M. (1993), *Schemata bei der Pluralbildung im Deutschen: Versuch einer kognitiven Morphologie*, Vol. 47, G. Narr.

Köpcke, K.-M. (1998), 'The acquisition of plural marking in English and German revisited: schemata versus rules', *Journal of Child Language* **25**(2), 293–319.
URL: *https://www.cambridge.org/core/product/identifier/S0305000998003407/type/journal*$_{a}rticle$

Köpcke, K.-M., Schimke, S. and Wecker, V. (2021), 'Processing of German noun plurals: Evidence for first- and second-order schemata', *Word Structure* **14**(1), 1–24.
URL: *https://www.euppublishing.com/doi/10.3366/word.2021.0173*

Köpcke, K.-M. and Wecker, V. (2017), 'Source- and product-oriented strategies in L2 acquisition of plural marking in German', *Morphology* **27**(1), 77–103.
URL: *http://link.springer.com/10.1007/s11525-016-9292-5*

Köpcke, K.-M. and Zubin, D. A. (1984), *Sechs Prinzipien für die Genuszuweisung im Deutschen: ein Beitrag zur natürlichen Klassifikation*.

Lin, J. (1991), 'Divergence measures based on the Shannon entropy', *IEEE Transactions on Information Theory* **37**(1), 145–151.
URL: *http://ieeexplore.ieee.org/document/61115/*

Liu, L. and Hulden, M. (2021), 'Can a Transformer Pass the Wug Test? Tuning Copying Bias

in Neural Morphological Inflection Models', *arXiv:2104.06483 [cs]* . arXiv: 2104.06483.
URL: *http://arxiv.org/abs/2104.06483*

MacWhinney, B. and Snow, C. (1985), 'The child language data exchange system*', *Journal of Child Language* **12**(2), 271–295. Publisher: Cambridge University Press.
URL: *http://www.cambridge.org/core/journals/journal-of-child-language/article/child-language-data-exchange-system/914F419D804ECA0C726F0C08E7E3C094*

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B. and Fedorenko, E. (2023), 'Dissociating language and thought in large language models: a cognitive perspective'. arXiv:2301.06627 [cs].
URL: *http://arxiv.org/abs/2301.06627*

Mahowald, K., Kachergis, G. and Frank, M. C. (2020), 'What counts as an exemplar model, anyway? A commentary on Ambridge (2020)'.
URL: *https://psyarxiv.com/ut86f/*

Malouf, R. (2017), 'Abstractive morphological learning with a recurrent neural network', *Morphology* **27**(4), 431–458.
URL: *http://link.springer.com/10.1007/s11525-017-9307-x*

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R. and Pinker, S. (1995), 'German inflection: The exception that proves the rule', *Cognitive psychology* **29**(3), 189–256.

Marr, D. (1982), *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, MIT Press. Google-Books-ID: D8XxCwAAQBAJ.

McCarthy, A. D., Vylomova, E., Wu, S., Malaviya, C., Wolf-Sonkin, L., Nicolai, G., Kirov, C., Silfverberg, M., Mielke, S. J., Heinz, J., Cotterell, R. and Hulden, M. (2019), The SIGMORPHON 2019 Shared Task: Morphological Analysis in Context and Cross-Lingual Transfer for Inflection, *in* 'Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology', Association for Computational Linguistics, Florence, Italy, pp. 229–244.
URL: *https://www.aclweb.org/anthology/W19-4226*

McCoy, R. T., Frank, R. and Linzen, T. (2020), ‘Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks’, *Transactions of the Association for Computational Linguistics* **8**, 125–140.
URL: *https://doi.org/10.1162/tacl*$_{a0}0304$

McCurdy, K., Goldwater, S. and Lopez, A. (2020), Inflecting When There’s No Majority: Limitations of Encoder-Decoder Neural Networks as Cognitive Models for German Plurals, *in* ‘Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics’, Association for Computational Linguistics, Online, pp. 1745–1756.
URL: *https://www.aclweb.org/anthology/2020.acl-main.159*

McCurdy, K., Goldwater, S. and Lopez, A. (2022), ‘Regularization or lexical probability-matching? How German speakers generalize plural morphology’, *Proceedings of the Annual Meeting of the Cognitive Science Society* **44**(44).
URL: *https://escholarship.org/uc/item/61v0r6f1*

McCurdy, K., Lopez, A. and Goldwater, S. (2020*a*), Conditioning, but on Which Distribution? Grammatical Gender in German Plural Inflection, *in* ‘Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics’, Association for Computational Linguistics, Online, pp. 59–65.
URL: *https://www.aclweb.org/anthology/2020.cmcl-1.8*

McCurdy, K., Lopez, A. and Goldwater, S. (2020*b*), Modeling grammatical gender and plural inflection in German, *in* ‘Proceedings of the 26 Architectures and Mechanisms for Language Processing Conference (AMLaP)’, Universität Potsdam.
URL: *https://amlap2020.github.io/a/261.pdf*

McCurdy, K. and Serbetçi, O. (2017), Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings, *in* ‘Presented at WiNLP (Women in Natural Language Processing)’, Vancouver, Canada.
URL: *http://www.winlp.org/wp-content/uploads/2017/final*$_{p}apers_{2}017/46_{p}aper.pdf$

Milin, P., Keuleers, E. and Đurđević, D. (2011), ‘Allomorphic responses in Serbian pseudo-nouns as a result of analogical learning’, *Acta Linguistica Hungarica* **58**(1), 65–84.
URL: *http://www.akademiai.com/doi/abs/10.1556/ALing.58.2011.1-2.4*

Molloy, E. (2018), ‘You say Boykotts, I say Boykotte, let’s call the whole thing off: Exploring the profitability of {-s} and its place in the Modern Standard German plural inflection system’, p. 24.

Mugdan, J. (1977), *Flexionsmorphologie und Psycholinguistik: Unters. zu sprachl. Regeln u. ihrer Beherrschung durch Aphatiker, Kinder u. Ausländer, am Beispiel d. dt. Substantivdeklination*, Vol. 82, TBL-Verlag Narr.

Nakisa, R. C. and Hahn, U. (1996), ‘Where Defaults Don’t Help: the Case of the German Plural System’, *arXiv:cmp-lg/9605020* . arXiv: cmp-lg/9605020.
URL: *http://arxiv.org/abs/cmp-lg/9605020*

Newport, E. L. (2016), ‘Statistical language learning: computational, maturational, and linguistic constraints’, *Language and Cognition* **8**(3), 447–461.
URL: *https://www.cambridge.org/core/product/identifier/S186698081600020X/type/journal*$_{a}rticle$

Newport, E. L. (2020), ‘Children and Adults as Language Learners: Rules, Variation, and Maturational Change’, *Topics in Cognitive Science* **12**(1), 153–169. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12416.
URL: *https://www.onlinelibrary.wiley.com/doi/abs/10.1111/tops.12416*

Nieder, J., van de Vijver, R. and Tomaschek, F. (2022), ‘“All mimsy were the borogoves” – a discriminative learning model of morphological knowledge in pseudo-word inflection’, *Language, Cognition and Neuroscience* pp. 1–18.
URL: *https://www.tandfonline.com/doi/full/10.1080/23273798.2022.2127805*

Nieder, J., van de Vijver, R. and Mitterer, H. (2020), ‘Knowledge of Maltese singular–plural mappings: Analogy explains it best’, *Morphology* .
URL: *http://link.springer.com/10.1007/s11525-020-09353-7*

Nosofsky, R. M. (1988), ‘Exemplar-based accounts of relations between classification, recognition, and typicality’, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **14**(4), 700–708.

O’Hara, C. (2020), ‘Frequency Matching Behavior in On-line MaxEnt Learners’, *Proceedings*

of the Society for Computation in Linguistics **3**(1), 463–465.
URL: *https://scholarworks.umass.edu/scil/vol3/iss1/51/*

O'Donnell, T. J., Goodman, N. D. and Tenenbaum, J. B. (2009), Fragment Grammars: Exploring Computation and Reuse in Language, Technical report, MIT.

Perfors, A. (2012), 'When do memory limitations lead to regularization? An experimental and computational investigation', *Journal of Memory and Language* **67**(4), 486–506.
URL: *https://www.sciencedirect.com/science/article/pii/S0749596X12000800*

Perfors, A. (2016), 'Adult Regularization of Inconsistent Input Depends on Pragmatic Factors', *Language Learning and Development* **12**(2), 138–155.
URL: *http://www.tandfonline.com/doi/full/10.1080/15475441.2015.1052449*

Perfors, A., Tenenbaum, J. B. and Wonnacott, E. (2010), 'Variability, negative evidence, and the acquisition of verb argument constructions*', *Journal of Child Language* **37**(3), 607–642. Publisher: Cambridge University Press.
URL: *https://www.cambridge.org/core/journals/journal-of-child-language/article/variability-negative-evidence-and-the-acquisition-of-verb-argument-constructions/D62EDBFF5A8F1ACC821451FEAD3C88FB*

Pierrehumbert, J. B. (2022), More than seventy years of probabilistic phonology, *in* B. E. Dresher and H. van der Hulst, eds, 'The Oxford History of Phonology', 1 edn, Oxford University PressOxford, pp. 639–655.
URL: *https://academic.oup.com/book/41916/chapter/354812559*

Pimentel, T., Ryskina, M., Mielke, S. J., Wu, S., Chodroff, E., Leonard, B., Nicolai, G., Ghanggo Ate, Y., Khalifa, S., Habash, N., El-Khaissi, C., Goldman, O., Gasser, M., Lane, W., Coler, M., Oncevay, A., Montoya Samame, J. R., Silva Villegas, G. C., Ek, A., Bernardy, J.-P., Shcherbakov, A., Bayyr-ool, A., Sheifer, K., Ganieva, S., Plugaryov, M., Klyachko, E., Salehi, A., Krizhanovsky, A., Krizhanovsky, N., Vania, C., Ivanova, S., Salchak, A., Straughn, C., Liu, Z., Washington, J. N., Ataman, D., Kieraś, W., Woliński, M., Suhardijanto, T., Stoehr, N., Nuriah, Z., Ratan, S., Tyers, F. M., Ponti, E. M., Aiton, G., Hatcher, R. J., Prud'hommeaux, E., Kumar, R., Hulden, M., Barta, B., Lakatos, D., Szolnok, G., Ács,

J., Raj, M., Yarowsky, D., Cotterell, R., Ambridge, B. and Vylomova, E. (2021), SIGMORPHON 2021 Shared Task on Morphological Reinflection: Generalization Across Languages, *in* 'Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology', Association for Computational Linguistics, Online, pp. 229–259.
URL: *https://aclanthology.org/2021.sigmorphon-1.25*

Pinker, S. and Prince, A. (1988), 'On language and connectionism: Analysis of a parallel distributed processing model of language acquisition', *Cognition* **28**(1-2), 73–193.
URL: *http://linkinghub.elsevier.com/retrieve/pii/0010027788900327*

Pinker, S. and Ullman, M. T. (2002), 'The past and future of the past tense', *Trends in Cognitive Sciences* **6**(11), 456–463.
URL: *https://www.sciencedirect.com/science/article/pii/S1364661302019903*

Prasada, S. and Pinker, S. (1993), 'Generalisation of regular and irregular morphological patterns', *Language and Cognitive Processes* **8**(1), 1–56.
URL: *https://www.tandfonline.com/doi/full/10.1080/01690969308406948*

Prince, A. and Smolensky, P. (2004), *Optimality Theory: Constraint interaction in generative grammar*, John Wiley & Sons.

R Core Team (2019), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Ramscar, M. (2021), 'A discriminative account of the learning, representation and processing of inflection systems', *Language, Cognition and Neuroscience* **0**(0), 1–25. Publisher: Routledge _eprint: https://doi.org/10.1080/23273798.2021.2014062.
URL: *https://doi.org/10.1080/23273798.2021.2014062*

Ramscar, M., Dye, M. and McCauley, S. M. (2013), 'Error and expectation in language learning: The curious absence of mouses in adult speech', *Language* **89**(4), 760–793.
URL: *http://muse.jhu.edu/content/crossref/journals/language/v089/89.4.ramscar.html*

Rasin, E., Berger, I., Lan, N., Shefi, I. and Katzir, R. (2021), 'Approaching explanatory adequacy in phonology using Minimum Description Length', *Journal of Language Modelling* **9**(1), 17–66. Number: 1.
URL: *https://jlm.ipipan.waw.pl/index.php/JLM/article/view/266*

Razeghi, Y., Logan IV, R. L., Gardner, M. and Singh, S. (2022), 'Impact of Pretraining Term Frequencies on Few-Shot Reasoning'. arXiv:2202.07206 [cs].
URL: *http://arxiv.org/abs/2202.07206*

Reali, F. and Griffiths, T. L. (2009), 'The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning', *Cognition* **111**(3), 317–328.
URL: *https://www.sciencedirect.com/science/article/pii/S0010027709000432*

Reeder, P. A., Newport, E. L. and Aslin, R. N. (2017), 'Distributional learning of subcategories in an artificial grammar: Category generalization and subcategory restrictions', *Journal of Memory and Language* **97**, 17–29.
URL: *https://www.sciencedirect.com/science/article/pii/S0749596X17300529*

Reilly, J. and Kean, J. (2007), 'Formal Distinctiveness of High- and Low-Imageability Nouns: Analyses and Theoretical Implications', *Cognitive Science* **31**(1), 157–168. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1080/03640210709336988.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1080/03640210709336988*

Rissanen, J. (1983), 'A Universal Prior for Integers and Estimation by Minimum Description Length', *The Annals of Statistics* **11**(2), 416–431. Publisher: Institute of Mathematical Statistics.
URL: *https://projecteuclid.org/journals/annals-of-statistics/volume-11/issue-2/A-Universal-Prior-for-Integers-and-Estimation-by-Minimum-Description/10.1214/aos/1176346150.full*

Rosen, E. R. (2022), Modeling human-like morphological prediction, *in* 'Proceedings of the

Society for Computation in Linguistics', Vol. 5, pp. 133–142.
URL: *https://scholarworks.umass.edu/scil/vol5/iss1/12/*

Round, E., Beniamine, S., Esher, L., Spike, M., Mann, Stephan and Lindsay-Smith, E. (2022), Cognition and the stability of evolving complex morphology: an agent-based model, *in* 'The Evolution of Language: Proceedings of the Joint Conference on Language Evolution (JCoLE)', Joint Conference on Language Evolution (JCoLE), pp. 635–642.

Royle, P., Beritognolo, G. and Bergeron, E. (2012), 'Regularity, sub-regularity and irregularity in French acquisition', *Irregularity in morphology (and beyond). Berlin: Akademie Verlag* pp. 227–250.

Rumelhart, D. E. and McClelland, J. (1986), On Learning the Past Tenses of English Verbs, *in* 'Parallel distributed processing: Explorations in the microstructure of cognition', MIT Press, Cambridge, MA, pp. 216–271.

Rácz, P., Beckner, C., Hay, J. B. and Pierrehumbert, J. B. (2020), 'Morphological convergence as on-line lexical analogy', *Language* **96**(4), 735–770. Publisher: Linguistic Society of America.
URL: *https://muse.jhu.edu/pub/24/article/775364*

Saldana, C., Claidière, N., Fagot, J. and Smith, K. (2022), 'Probability matching is not the default decision making strategy in human and non-human primates', *Scientific Reports* **12**(1), 13092. Number: 1 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41598-022-16983-w*

Samara, A., Smith, K., Brown, H. and Wonnacott, E. (2017), 'Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation', *Cognitive Psychology* **94**, 85–114.
URL: *https://www.sciencedirect.com/science/article/pii/S0010028516301426*

Schuhmann, K. S. and Putnam, M. T. (2021), 'Relativized Prosodic Domains: A Late-Insertion Account of German Plurals', *Languages* **6**(3), 142. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.
URL: *https://www.mdpi.com/2226-471X/6/3/142*

Schütze, C. T. (2005), Thinking About What We Are Asking Speakers to Do, *in* H. v. Riemsdijk, H. v. d. Hulst, J. Koster, S. Kepser and M. Reis, eds, 'Linguistic Evidence', Vol. 85, Mouton de Gruyter, Berlin, New York, pp. 457–484. Series Title: Studies in Generative Grammar.
URL: *https://www.degruyter.com/view/books/9783110197549/9783110197549.457/9783110197549.457.xml*

Segbers, J. and Schroeder, S. (2017), 'How many words do children know? A corpus-based estimation of children's total vocabulary size', *Language Testing* **34**(3), 297–320.
URL: *http://journals.sagepub.com/doi/10.1177/0265532216641152*

Seidenberg, M. S. and Plaut, D. C. (2014), 'Quasiregularity and Its Discontents: The Legacy of the Past Tense Debate', *Cognitive Science* **38**(6), 1190–1228.
URL: *http://doi.wiley.com/10.1111/cogs.12147*

Seyfarth, S., Ackerman, F. and Malouf, R. (2014), 'Implicative organization facilitates morphological learning', *Annual Meeting of the Berkeley Linguistics Society* **40**, 480.
URL: *http://journals.linguisticsociety.org/proceedings/index.php/BLS/article/view/3154*

Shanks, D. R., Tunney, R. J. and McCarthy, J. D. (2002), 'A re-examination of probability matching and rational choice', *Journal of Behavioral Decision Making* **15**(3), 233–250. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.413.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.413*

Shannon, C. E. (1948), 'A mathematical theory of communication', *The Bell System Technical Journal* **27**(3), 379–423. Conference Name: The Bell System Technical Journal.

Skousen, R. (1989), *Analogical modeling of language / Royal Skousen.*, Kluwer Academic Publishers, Dordrecht ;. Publication Title: Analogical modeling of language.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K. and Wonnacott, E. (2017), 'Language learning, language use and the evolution of linguistic variation', *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**(1711), 20160051.
URL: *https://royalsocietypublishing.org/doi/10.1098/rstb.2016.0051*

Smith, K. and Wonnacott, E. (2010), 'Eliminating unpredictable variation through iterated

learning', *Cognition* **116**(3), 444–449.
URL: *https://www.sciencedirect.com/science/article/pii/S0010027710001320*

Smolensky, P. and Legendre, G. (2006), *The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture), Vol. 1*, MIT press.

Sonnenstuhl, I. and Huth, A. (2002), 'Processing and Representation of German -n Plurals: A Dual Mechanism Approach', *Brain and Language* **81**(1-3), 276–290.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0093934X01925246*

Soulos, P., Hu, E., McCurdy, K., Chen, Y., Fernandez, R., Smolensky, P. and Gao, J. (2023), Differentiable Tree Operations Promote Compositional Generalization, *in* 'Proceedings of the 40th International Conference on Machine Learning', Proceedings of Machine Learning Research, PMLR.

Spreng, B. (2004), 'Error patterns in the acquisition of German plural morphology: Evidence for the relevance of grammatical gender as a cue', *Toronto Working Papers in Linguistics* **23**.
URL: *https://twpl.library.utoronto.ca/index.php/twpl/article/view/6211*

Stemberger, J. P. (2006), Proper Names in Slovene: Implications for defaults in inflectional morphology, *in* 'Proceedings of the Annual Meeting of the Cognitive Science Society', Vol. 28. Issue: 28.

Sutskever, I., Vinyals, O. and Le, Q. V. (2014), Sequence to Sequence Learning with Neural Networks, *in* 'Advances in Neural Information Processing Systems', pp. 3104–3112.

Sylak-Glassman, J., Kirov, C., Yarowsky, D. and Que, R. (2015), A Language-Independent Feature Schema for Inflectional Morphology, *in* 'Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)', Association for Computational Linguistics, Beijing, China, pp. 674–680.
URL: *http://aclweb.org/anthology/P15-2111*

Trommer, J. (2020), 'The subsegmental structure of German plural allomorphy', *Natural Lan-*

guage & Linguistic Theory .

URL: *http://link.springer.com/10.1007/s11049-020-09479-7*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, . and Polosukhin, I. (2017), Attention is all you need, *in* 'Proceedings of the 31st International Conference on Neural Information Processing Systems', NIPS'17, Curran Associates Inc., Red Hook, NY, USA, pp. 6000–6010.

Veríssimo, J. and Clahsen, H. (2014), 'Variables and similarity in linguistic generalization: Evidence from inflectional classes in Portuguese', *Journal of Memory and Language* **76**, 61–79.

URL: *https://www.sciencedirect.com/science/article/pii/S0749596X1400059X*

Vulkan, N. (2000), 'An Economist's Perspective on Probability Matching', *Journal of Economic Surveys* **14**(1), 101–118. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-6419.00106.

URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-6419.00106*

Wegener, H. (1994), Variation in the Acquisition of German Plural Morphology by Second Language Learners, *in* 'How tolerant is universal grammar? Essays on language learnability and language variation', number 309 *in* 'Linguistische Arbeiten', Max Niemeyer, Tübingen, pp. 267–294.

Weir, M. W. (1964), 'Developmental changes in problem-solving strategies', *Psychological Review* **71**, 473–490. Place: US Publisher: American Psychological Association.

Wiese, R. (1996), *The phonology of German*, Oxford University Press on Demand.

Wiese, R. (1999), 'On default rules and other rules', *Behavioral and Brain Sciences* **22**(6), 1043–1044.

URL: *https://www.cambridge.org/core/product/identifier/S0140525X99532226/type/journal*$_article$

Wilder, C. and Gärtner, H.-M. (1997), 'Zur Rolle von Ökonomieprinzipien in der Generativen Linguistik', *Zeitschrift für Sprachwissenschaft* **16**(1-2), 227–264. Publisher: De Gruyter.

URL: *https://www.degruyter.com/document/doi/10.1515/zfsw.1997.16.1-2.227/html*

Wilder, C., Gärtner, H.-M. and Bierwisch, M. (1996), *The Role of Economy Principles in Linguistic Theory*, Walter de Gruyter GmbH & Co KG. Google-Books-ID: Lg90DwAAQBAJ.

Williams, A., Pimentel, T., Blix, H., McCarthy, A. D., Chodroff, E. and Cotterell, R. (2020), Predicting Declension Class from Form and Meaning, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 6682–6695.
URL: *https://www.aclweb.org/anthology/2020.acl-main.597*

Wonnacott, E., Newport, E. L. and Tanenhaus, M. K. (2008), 'Acquiring and processing verb argument structure: Distributional learning in a miniature language', *Cognitive Psychology* **56**(3), 165–209.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S001002850700031X*

Wu, S., Cotterell, R. and O'Donnell, T. (2019), Morphological Irregularity Correlates with Frequency, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 5117–5126.
URL: *https://www.aclweb.org/anthology/P19-1505*

Wulf, D. J. (2002), Applying analogical modeling to the German plural, *in* 'Analogical modeling', John Benjamins, Amsterdam, pp. 109–122.
URL: *http://ebookcentral.proquest.com/lib/ed/detail.action?docID=622222*

Wurzel, W. U. (1970), *Studien zur deutschen Lautstruktur*, Vol. 8, Akademie-Verlag.

Yang, C. (2017), How to wake up irregular (and speechless), *in* 'On looking into words (and beyond)', Language Science Press Berlin, pp. 211–233. Publisher:.

Yang, C. D. (2016), *The price of linguistic productivity : how children learn to break the rules of language*, The MIT Press, Cambridge, Massachusetts.

Zaretsky, E. and Lange, B. P. (2016), No matter how hard we try: Still no default plural marker in nonce nouns in Modern High German, *in* 'A blend of MaLT: selected contributions from the Methods and Linguistic Theories Symposium 2015', number Band 15 *in* 'Bamberger Beiträge zur Linguistik', University of Bamberg Press, Bamberg, pp. 153–178.

Zaretsky, E., Lange, B. P., Euler, H. A. and Neumann, K. (2013), 'Acquisition of German pluralization rules in monolingual and multilingual children', *Studies in Second Language Learning and Teaching* **3**(4), 551.
URL: *http://pressto.amu.edu.pl/index.php/ssllt/article/view/5029*

Zwicky, A. M. (1985), How to describe inflection, *in* 'Annual Meeting of the Berkeley Linguistics Society', Vol. 11, pp. 372–386.

Zwicky, A. M. (1986), The general case: Basic form versus default form, *in* 'Annual Meeting of the Berkeley Linguistics Society', Vol. 12, pp. 305–314.

Zymet, J. (2019), 'Learning a Frequency-Matching Grammar together with Lexical Idiosyncrasy: MaxEnt versus Hierarchical Regression', *Proceedings of the Annual Meetings on Phonology* **7**.
URL: *http://journals.linguisticsociety.org/proceedings/index.php/amphonology/article/view/4495*