[t]

Neural Networks Don't Learn Default Rules for German Plurals, But That's Okay, Neither Do Germans

Kate McCurdy

Master of Science by Research

Centre for Doctoral Training in Data Science

School of Informatics

University of Edinburgh

2019

Abstract

Can artificial neural networks learn to represent inflectional morphology and generalize to new words as human speakers do? Some linguists have argued that the German number system cannot be modeled without rule-based symbolic computation, because the 'default' plural marker, /-s/, is also the least frequent; they claim that speaker preferences for /-s/ in *elsewhere* conditions, such as novel and phonologically atypical nouns, require representation of linguistic rules and thus cannot be learned from data alone.

We present a new dataset of German speakers' production and rating of plural forms for novel nouns, and note that the results provide at best weak support the claimed 'default' status for /-s/, reducing its potential challenge for neural models. Nonetheless, we observe that neural encoder-decoder models, while broadly successful on this 'wug' task, show distinctive failure modes suggesting they do not generalize in quite the same manner as human speakers.

Acknowledgements

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

And you — you know who you are.

Table of Contents

Chapter 1

Introduction

Morphology, the "study of systematic covariation in the form and meaning of words" (Haspelmath and Sims, 2010, 2), has historically been the site of vigorous debate on the capacity of models, neural models in particular, to model (human) speaker behavior, and hence attempt to make claims about speaker cognition. In 1986, Rumelhart and McClelland described a neural network model which learned to map English verbs in the present tense to their past tense forms. Importantly, the network handled both *regular* verbs with past tense inflection formed systematically by adding the suffix /-d/ (e.g. *jumped*), and *irregular* verbs whose past tense form was partially or wholly unpredictable given the present tense input (e.g. *ran*). The authors suggested their model provided "an alternative [...] to the implicit knowledge of rules" (1986, 218) which until then undergirded not only past tense inflection (Halle, 1973), but linguistic phenomena overall; in fact, their model was the first to depict a process of linguistic generalization that was not characterized by application of a specific symbolic rule (Seidenberg and Plaut, 2014). Rumelhart and McClelland's claims sparked controversy and were swiftly and vehemently contested. Most notably, Pinker and Prince (1988) highlighted many empirical inadequacies of the RM model, and argued that these failings were inherent to "central features of connectionist ideology" and would therefore persist in any neural network model lacking a symbolic processing component. One of the core theoretical deficiencies they identified in connectionist (or 'pattern associator') models relative to symbolic processing was the inability to represent a *default* structure: whereas a rule such as stem + /-d/ can cover all possible cases irrespective of the form of the input (i.e. the stem), a neural network can only learn subregularities based on exemplars from observed inputs — in principle, it can't learn to represent a default behavior which applies in the absence of associated patterns (the *elsewhere* condition).

Marcus et al. (1995) reasserted the importance of default rule representation in the context of German number inflection, and identified German's plural class system as a crucial test case for connectionist modeling: while the RM model might successfully learn English past tense inflection due to its historical confound of frequency (95% of verbs take past /-d/) and regularity (/-d/ is the productive, default past tense marker), German plurals are claimed to be a *minority-default* system in which these two factors are decoupled. Two core claims from that paper motivate the key research questions of this thesis:

1. the German plural marker /-s/, despite constituting a numerical minority (applying to 4% of noun types, less frequent than 4 other plural suffixes; c.f. Table 2.1), nonetheless constitutes the *default* plural class which applies in the elsewhere condition (a claim they support with evidence from speaker judgments of novel nouns); and

2. neural networks should struggle to generalize /-s/ correctly, as its low frequency and highly variable distribution (a result of its 'elsewhere' application) cannot easily be learned by extrapolation from observed patterns in the data (a claim they do not test directly).

In the years since the paper was published, a number of developments have called these claims into question.

Claim 1, for /-s/ as the default plural marker, has frequently been challenged with behavioral evidence for the productivity of other plural classes (e.g. Köpcke, 1998; Wiese, 1999; Yang, 2016); however, the strongest counterargument has come from a recent large-scale production study using the same stimuli, which failed to reproduce the preference for /-s/ found in the original experiment (Zaretsky and Lange, 2016). Even if the distribution of the /-s/ plural marker over German nouns can best be characterized as a reflection of the elsewhere condition, it is unclear whether this analysis meaningfully predicts speakers' propensity to generalize the /-s/ class to novel nouns.

Claim 2 has also been challenged, as a new generation of powerful neural networks points toward the expanded capacity of connectionist models to process and represent natural language (Pater, 2019). Kirov and Cotterell (2018) demonstrate that modern encoder-decoder models overcome many of Pinker and Prince's criticisms and successfully learn English past tense inflection, even producing human-like errors by generalizing the regular /-d/ suffix to a held-out test set. They argue that neural networks are worth reconsideration as cognitive models for language processing, sidestepping

the 'rules' debate to instead posit two evaluation criteria: "(i) Does the learner induce the full set of correct generalizations about the data? Given a range of novel inputs, to what extent does it apply the correct transformations to them? (ii) Does the behavior of the learner mimic humans? Are the errors human-like?" (2018, 2). The success of encoder-decoder models on English past tense inflection is certainly noteworthy, but Marcus et al.'s criticism of the frequency-regularity confound still stands: a network that generalizes the most frequent suffix to novel inputs may not be able to generalize lower-frequency classes, such as those of the German plural system.

1.1 Research questions

In light of the developments outlined above, this thesis project revisits Marcus et al.'s core claims and investigates them as related research questions.

RQ1. Do German speakers treat /-s/ as the default plural for novel nouns? Comparing the behavior of modern neural networks and human speakers on German plurals requires a detailed overview of speaker behavior, but currently no such data is publicly available. We collect new German speaker data on the same stimuli used by Marcus et al. (1995), gathering production and rating judgments. The results show a strong quantitative preference for /-e/ and /-en/ over /-s/, and evidence consistent with both /-en/ and /-s/ as productive classes for these stimuli.

RQ2. Do modern neural networks learn the correct generalizations for the German plural system? Following the architecture used by Kirov and Cotterell, we train an encoder-decoder to map German nouns in the singular to their plural forms. The model's outputs are evaluated on i) a held-out test set of existing German nouns, and ii) the novel noun stimuli with speaker judgments from RQ1. On the test set, the model achieves high accuracy overall, but struggles to predict the /-s/ class accurately, a finding consistent with Marcus et al.'s predictions. On the novel words, the model's predictions broadly approximate speaker behavior (both align on /-e/ as the most frequent class), but show inconsistent behavior in other respects. Most notably, the model fails to generalize /-en/ to the extent speakers do, suggesting it does not provide a convincing analogue to human learners. While this finding is not conclusive — the model lacked access to noun class information via determiners, for example, which could improve performance substantially — it nonetheless highlights limitations which may restrict neural networks' utility as models of cognition. For all of deep learning's power, some of the old criticisms retain their validity.

1.2 Thesis outline

The remainder of this thesis is organized as follows:

- Chapter 2, **Background**, gives an overview of the German plural system and introduces the *wug test*, a core methodology for behavioral studies of inflectional morphology for both humans and neural networks.
- Chapter 3, **German Speaker Data**, presents the investigation and analysis of RQ1: wug testing of production and rating via an online survey.
- Chapter 4, **Encoder-Decoder Model**, covers the first part of RQ2: the neural network implementation, and results on the held-out test set.
- Chapter 5, **Comparison and Analysis**, covers the second part of RQ2, with a detailed comparison of speaker and model behavior on the wug test stimuli.

Chapter 2

Background

2.1 Modeling morphology

Morphological inflection has been the site of vigorous debate concerning linguistic representation and cognition for decades, beginning with the influential attempt by Rumelhart and McClelland (1986) to model both regular and irregular past tense inflection of English verbs with a single neural network, and the famously scathing rebuttal of Pinker and Prince (1988). The core point of contention is the role of rule-based symbolic computation in modeling morphological regularity. While linguists and cognitive scientists in favor of 'single-route' models argue that a single learning mechanism can account for regular and irregular processes (usually on the basis of analogical similarity across patterns or schemas, e.g. Bybee, 1995; Köpcke, 1988; Seidenberg and McClelland, 1989; Plunkett, 1993; Blything et al., 2018)[1], proponents of the 'dual-route' model posit that regular morphological processes are necessarily computed by categorical symbolic rules, while irregular morphological forms are stored in memory and may only sporadically extend to new words via analogical processes (Prasada and Pinker, 1993; Pinker, 1999; Clahsen, 1999b). At the heart of the debate is the question of generalization: how do human speakers apply inflection to unseen words — a task which they must do regularly, in the face of a constantly changing vocabulary — and what does a computational model require to replicate this behavior?

[1] Although rule-based single-route models have also been proposed (Albright and Hayes, 2003; Yang, 2016).

2.1.1 Wug tests

The most widely used method for experimental investigation of speakers' inflectional behavior is known as the 'wug' test (Berko, 1958). To conduct a wug test, researchers make up a series of 'nonce' words (i.e. words that are phonologically plausible but do not exist in the language, such as *spling* or *dize* in English; Albright and Hayes, 2003), then present experimental participants with the task of inflecting the novel words, e.g. producing the plural form of a singular word. The core assumption underlying the wug test is that speakers' inflection of unseen words must reflect the *productive* morphological processes of the language, as non-productive processes should not be generalized to new entries to the lexicon. The name 'wug' comes from one of the nonce words used in the first application of this method: Berko (1958) demonstrated that young English-speaking children successfully extend the /-z/ allomorph to produce the plural form *wugs*, indicating implicit knowledge of English grammatical processes. The method has since been extended to investigate different properties of linguistic variation across many languages (Kawahara, 2016).

Albright and Hayes (2003) used wug testing to compare the predictions of dual-route and single-route models of morphology against experimental data from human speakers. They presented participants with nonce verbs such as *spling*; speakers were prompted to produce the verb in the past tense, and to rate the acceptability of two or three past tense forms of verb, including the regular form (e.g. *splinged*) and one or two irregular forms developed by the authors (e.g. *splung*). Albright and Hayes found evidence for gradience in acceptability of regular forms, which is not predicted under categorical rule application in the dual-route model; however, they also found that a single-route analogical model based on phonological similarity relied too heavily on exemplars and failed to capture abstract patterns, leading the authors to prefer a stochastic rule-based learner. This thesis follows their experimental approach of collecting production and rating data from wug tests with speakers, and using the results to evaluate computational models of morphology.

2.1.2 Wug testing for neural networks

Over the past decade, advances in artificial neural networks have led to high performance on many natural language tasks, raising the possibility that neural approaches could productively inform linguistic inquiry (Pater, 2019). Recent work by Kirov and Cotterell (2018, henceforth K&C) revisits the debate surrounding the English past

tense. They use an encoder-decoder neural architecture trained on phoneme-level representations of verb stems and their past tense inflected forms, mirroring the supervised task modeled by Rumelhart and McClelland. The authors argue that their work overcomes the criticisms of Pinker and Prince (1988): the model can represent ordered phoneme sequences of variable length[2], and achieves near-ceiling accuracy on a held-out test set by correctly generalizing the regular /-d/ form. They also test the model on the nonce words evaluated by Albright and Hayes (2003), and report that the neural encoder-decoder scores yield a higher correlation with human speakers' production probabilities relative to the predictions of Albright and Hayes' rule-based learning model. While more recent work by Corkery et al. (2019) find that K&C's proposed model displays highly variable performance on the English nonce word past tense inflection task, calling the generalizability of this result into question, K&C nonetheless show that contemporary neural networks appear far better equipped to handle the task of morphological inflection compared to previous-generation models.

While K&C's work demonstrates the capacity of modern neural networks to model both regular and irregular English inflectional morphology, the core confound identified by Marcus et al. (1995) remains relevant: the regular suffix /-d/ is also by far the most frequent mechanism for past tense realization in English (both by word type frequency and by token frequency). This raises the possibly that neural models could be simply learning to produce high-frequency mappings in novel contexts, and are at risk of failing to generalize correctly when frequency is not the appropriate cue. If Marcus et al.'s analysis of the plural suffix /-s/ is correct, then plural inflection in German represents a suitable test case to evaluate whether modern neural models indeed learn human-like behavior with respect to inflectional morphology.

2.2 German plurals

The German plural system comprises five suffixes: /-e/, /-er/, /-ø/[3], /-en/, and /-s/. The first three can optionally combine with an umlaut over the root vowel[4], yielding eight potential plural markers (Clahsen et al., 1992). Examples in all forms are shown in Table 2.1. Each plural suffix is also shown with its type frequency (counting each

[2]Rumelhart and McClelland used fixed-length transformations of phonemic sequences termed 'Wickelphones,' which, as Pinker and Prince noted, lacked the capacity to represent core aspects of phonology such as sequential ordering.

[3]/-ø/ refers to the so-called "zero plural", and is indicated as "zero" on all figures in this paper.

[4]As umlaut is a process which fronts a back vowel, only roots with back vowels (e.g. *Dach*, *Fuss*) can take an umlaut; words with front vowels (e.g. *Kind*, *Bett*) are inherently excluded.

Suffix	Singular	Plural	Type	Token
/-(e)n/	die Strasse	die Strassen	48%	45%
/-e/	der Hund	die Hunde	27%	21%
	die Kuh	die Kühe		
/-ø/	der Daumen	die Daumen	17%	29%
	die Mutter	die Mütter		
/-er/	das Kind	die Kinder	4%	3%
	der Wald	die Wälder		
/-s/	das Auto	die Autos	4%	2%

Table 2.1: German plural system, ordered by CELEX type frequency (Sonnenstuhl and Huth, 2002).

word type only once, how many types take this plural?) and token frequency (how often do words with this plural suffix appear in the corpus overall?).

Like grammatical gender, the plural class is a lexical feature of a noun[5]. Compound nouns such as *das Kleinkind / die Kleinkinder* (the small child / the small children) inherit the grammatical gender (*das*; neuter) and plural class (/-er/) of the head noun *das Kind / die Kinder* (the child / the children). Derived nouns such as *die Reduzierung* (the reduction) also take the grammatical gender and plural class of the derivational suffix (here *-ung*, which takes the feminine article *die* and the plural suffix /-en/). Clearly, these word-formation processes contribute to statistical associations between plural class, grammatical gender, and phonological structure, especially a word's final syllable. Indeed, virtually all scholarly analyses of the German plural system agree that grammatical gender and lexical phonology interact with plural class, both for morphologically complex nouns (e.g. compounds, derived forms) and non-derived nouns (Wiese, 1996; Bittner, 2000; Laaha, 2011; Yang, 2016). However, the interpretation of these associations — whether they constitute productive forms of generalization, or historical relics of formerly productive processes visible on "irregular" nouns — has been an area of contestation.

[5]Although it can also be determined by certain syntactic or semantic contexts; see later this section.

2.2.1 Debating the distribution and productivity of plural classes

The dual-route analysis of German noun inflection classifies /-s/ as the *default* plural, and hence the only *regular*, productive plural suffix despite its low frequency, while the other plural classes all represent irregular inflection (Clahsen, 1999b). Marcus et al. (1995) argue that its propensity to appear in a range of *elsewhere conditions* demonstrate the status of /-s/ as the default plural marker. Examples include the requirement for /-s/ to appear on proper names (e.g. *der Bader / die Bader* 'the barber / the barbers' but *meine Freunden, die Baders* 'my friends, the Barbers'), acronyms, and truncated and quoted nouns (e.g. *der Asi / die Asis*, short for *Asozialer* 'antisocial person'). In addition, /-s/ tends to be the plural class for recent borrowings from other languages, and children reportedly overregularize /-s/ by extending it to novel nouns (Clahsen 1999b; Clahsen et al. 1992; although see Köpcke 1998). As these various environments do not easily fit into any coherent characterization, they are said to illustrate applications of a rule-driven default. Because /-s/ is a minority class in terms of type and token frequency, this analysis describes a *minority-default* system of number inflection.

Some researchers agree with the analysis of /-s/ as default (e.g. Janda, 1990); however, many others emphasize the evidence for productivity of other plural classes (e.g. Köpcke, 1988; Yang, 2016). Wiese (1996) and Wunderlich (1999) analyze /-(e)n/ as the default plural for feminine nouns, in keeping with acquisition studies that find children generalize /-(e)n/ (Köpcke, 1998; Elsen, 2002; Zaretsky et al., 2013). Other add /-e/ as a productive suffix (Dressler, 1999; Pulvermüller, 1999). Indefrey (1999, 1025) argues that /-(e)n/ and /-e/ are "regular and productive allomorphs with gender-dependent application domains", noting that /-e/ and /-en/ are extended in elsewhere conditions where /-s/ is blocked for phonological reasons, e.g. letters (*"X"e*), acronyms (*die MAZen, Magnetaufzeichnungen*, 'magnetic recordings'), and product names (*Mercedesse*, 'Mercedes cars'), although Clahsen (1999a) argues that these forms reflect independent prosodic constraints (c.f. Wiese, 1996) rather than default plural inflection. Bybee (1995) argues that while /-s/ *can* operate as a default plural, it also shows generalization patterns influenced by phonological similarity (such as the observation that nouns ending in full vowels such as *a, o, i* take /-s/; Elgersma and Houseman 1999); hence its behavior is consistent with acting both as default 'emergency plural' and as a lexically-influenced schema (Köpcke, 1988). Stemberger (1999, 1041) summarizes the evidence thusly: "/-s/ does not act like a default, so its low frequency is not a problem, but it does not act like other irregular patterns either."

Suffix	Feminine (*die*)	Masculine (*der*)	Neuter (*das*)
/-(e)n/	3475 (745)	643 (118)	75 (39)
/-e/	38 (35)	1676 (615)	1378 (592)
+uml	75 (20)	703 (140)	3 (2)
/-ø/	34 (26)	1470 (279)	553 (195)
+uml	4 (1)	93 (24)	6 (0)
/-er/	0	4 (0)	86 (19)
+uml	1 (1)	47 (11)	191 (42)
/-s/	36 (21)	210 (58)	200 (38)
other	25 (6)	83 (26)	132 (28)

Table 2.2: Distribution of plural class by gender in nouns from the Unimorph German dataset (Kirov et al., 2016). Parentheses indicate subset of nouns with grammatical gender determined by heuristic rather than exact lookup.

Computational modeling: Nakisa and Hahn (1996) simplified the task of German number inflection from transduction (producing the correct output form for a given input, e.g. *Kind* → *Kinder*) to classification (predicting the plural class label for a given input, e.g. *Kind* → */-er/*), and evaluated the performance of several computational models with single- and dual-route architectures on this task. They found that single-route models performed very competitively: a nearest-neighbor classifier based on phonological similarity achieved 71% accuracy on predicting the plural class for a held-out test set, and a three-layer feed-forward neural network achieved 83% accuracy. They also found that adding a rule component for /-s/ plurals, as advocated by the dual-route analysis, did not increase accuracy over the single-route neural model[6]. In follow-up work, Hahn and Nakisa (2000) further experimented with hybrid models learning a threshold of certainty for when to apply the /-s/ rule as opposed to pattern association, and found no performance gains. However, Pinker and Ullman (2002) dismiss the classification task as inadequate, arguing that rule application would nonetheless be needed to combine inputs and plural class labels to yield the output form.

Corpus distribution: Table 2.2 illustrates the distribution of plural class by grammatical gender[7] for nouns in the Unimorph German dataset (Kirov et al., 2016), which

[6]Nakisa and Hahn also simulated two artificial languages with different class distributions, and found the hybrid model did outperform single-route models when the regular class was homogeneously distributed across the input space.

[7]As UniMorph does not include grammatical gender for German, noun gender was obtained by merging with another dataset scraped from the same source, Wiktionary; the data is available at `github`.

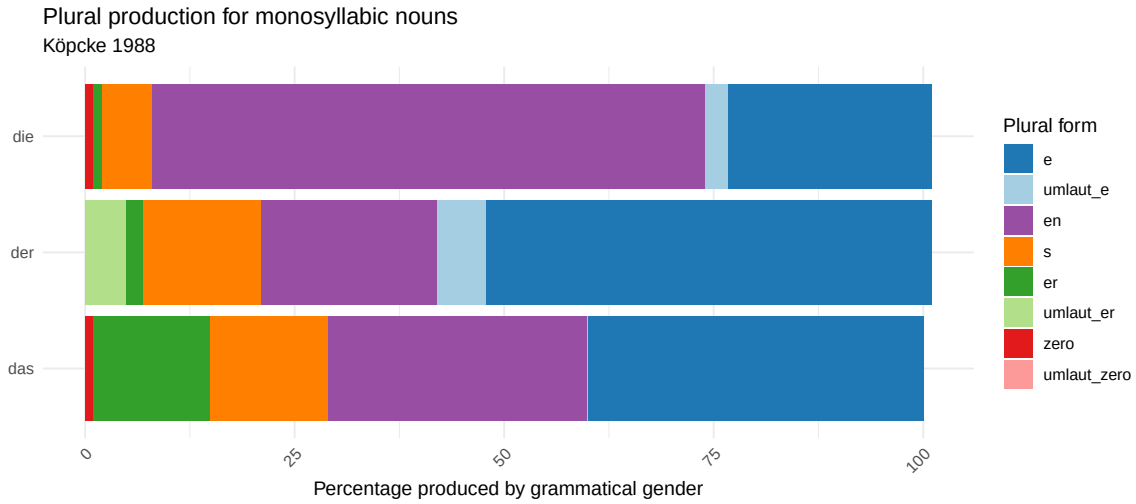

Figure 2.1: Plural class production for monosyllabic nouns by grammatical gender (Köpcke, 1988, Table 3, Class 5).

is the corpus used to train and evaluate the models described in Chapter 4. Wiese (1999, 1044) identifies three tendencies in gender-plural interaction, the first two of which are visible in the table. "(1) Feminine nouns predominantly take an /-n/ as plural affix" — true of 95% of feminine nouns in the corpus. "(2) Within the group of non-feminines, plural forms represented by the /-e/ suffix are found as well as plurals marked with the /-er/, but the latter plural is in a clear minority" — of masculine and neuter nouns together, 50% take /-e/ (including umlaut), but only 4% take /-er/ (including umlaut).[8] Another notable aspect of the table is the distribution of /-s/ across grammatical gender. Though /-s/ is clearly more prevalent in the non-feminine categories, it still appears fairly evenly across grammatical genders relative to other plural classes — an observation consistent with a widely-dispersed 'elsewhere' distribution.

2.2.2 Wug tests for German plurals

Köpcke (1988) conducted the first study to probe German speaker knowledge of plural formation through wug testing, using an auditory paradigm in which participants heard a nonce word, then spoke its plural aloud. His findings supported several previous theoretical observations: speakers generally produced the expected plural class

`com/gambolputty/german-nouns`. About 27% of UniMorph nouns were missing from the dataset; these nouns were heuristically assigned the gender of the noun which shared the longest suffix string. Counts of these 'guessed' nouns in each category are indicated with parentheses in Table 2.2.

[8] Wiese's third observation is that non-feminine nouns taking the /-ø/ plural all have a reduced final syllable, e.g. *Garten* (garden); monosyllabic words such as *Hund* (dog) never take a zero plural. This is relevant to the experimental design later, but not to the discussion of the table.

Factor	Condition	Examples (English translation)
Context	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.
Word familiarity	Rhyme	*Bral, Klot, Pind, Kach, Spert, Mur*
	Non-rhyme	*Plaupf, Pleik, Bnöhk, Fnahf, Snauk, Pläk*

Table 2.3: Study design and example stimuli for M95

associated with specific derivational suffixes (e.g. the masculine suffix *-ling*, which takes /-e/, and feminine *-schaft*, which takes /-en/); speakers predominantly produced /-(e)n/ for nouns ending in schwa of any gender; speakers predominantly produced /-s/ for nouns ending in a full, i.e. non-schwa, vowel. In addition, the novel experimental paradigm showed differences in how plural classes were generalized relative to their expected frequency in a given environment (where 'environment' is defined in terms of grammatical gender and final syllable). Outside of obligatory environments such as derivational suffixes, /-en/ and /-s/ were overgeneralized to a greater extent than /-e/, and /-er/ and zero were undergeneralized. Köpcke analyzed the speaker data in light of cue salience and historical trends toward /-en/ as a plural marker (especially for feminine-class nouns), and interpreted the results as consistent with a product-oriented 'schema' model of plural formation. His results for monosyllabic nouns, the main category of interest for following experiments, are shown in Figure 2.1.

To support the postulated analysis of /-s/ as the regular, default plural class for German, Marcus et al. (1995, henceforth M95) gathered experimental evidence using wug tests. They developed a set of 24 monosyllabic nonce nouns, deliberately avoiding phonological forms with predictable plural classes (such as the generalizations described by Köpcke (1988)). All of the items were 'legitimate' German words in the sense of obeying phonotactic constraints, but they nonetheless comprised two phonological classes: 'familiar' or 'rhyme' words with one or more existing rhyming words in German (e.g. *Bral*, rhyming with *Tal-Täler, Mal-Male, Pfahl-Pfähle, Wahl-Wahlen*; *Spert*, rhyming with *Wert-Werte, Pferd-Pferde*), and 'unfamiliar' or 'non-rhyme' words (e.g. *Plaupf*, *Bnöhk*), on the hypothesis that non-rhymes, as phonologically atypical

words, should be more likely to fall into the 'elsewhere' condition and thus more likely to take the /-s/ plural. These words were assigned a semantic meaning in one of three different conditions (as a proper 'Name', 'Borrowing' from another language, or 'Root' German word, again assuming the more atypical conditions would be more likely to trigger /-s/) and inserted into sentence context which reflected these semantic cues. Words in the Root and Borrowing conditions were counterbalanced for grammatical gender (50% masculine and 50% feminine in each list) across subjects. Examples of context and nonce words are given in Table 2.3.

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 (or five, if umlaut was not possible for the root word in question). M95 report a significant interaction between phonological class (rhyme or non-rhyme) and ratings for /-s/ plural forms, going in the predicted direction: /-s/ appeared as the top-rated plural form for 2 out of 12 rhyme words, and 7 out of 12 non-rhyme words. No significant interaction with gender was found, in contrast to the results of Köpcke (1988) for monosyllabic nonce nouns. M95 report averaged ratings for each item and plural variant, enabling relatively fine-grained analysis of their results, as we present in Chapter 3.

The evidence for /-s/ as default plural presented by M95, along with studies on verb inflection and other methods (priming, neuroimaging), has been used to argue that a dual-route model of regular (rule-based) and irregular (memory-based) processing best accounts for German inflection (Clahsen, 1999b). However, a range of challenges have been raised to this analysis, including the charge of failure to replicate. In a large-scale follow-up study, Zaretsky and Lange (2016, henceforth Z&L) used the same nonce words to elicit written productions of plural forms from 585 participants. In addition to altering the task (production instead of rating), Z&L also eliminated the Context factor of the original study, presenting words in 'Root' form[9] counterbalanced into feminine and non-feminine[10] item groups for presentation. They found that the plural suffix /-e/ predominated for both non-rhymes (52% of productions, of which 6% with umlaut) and rhymes (58% of productions, of which 14% with umlaut); moreover, while /-s/ showed the anticipated preference for non-rhymes (15%, vs. 8% for rhymes), /-en/

[9]It is unclear from the study description whether words were presented in the Root-context sentences from the original M95 study, or simply as bare nouns in isolation. As the study material in appendices from M95 did not include these context sentences, and the authors of M95 note that presenting words in isolation is consistent with a 'Root' interpretation, the general principle of simplicity/least effort suggests that the words were presented in isolation, preceded by an article.

[10]Whether 'non-feminine' entails masculine or neuter is left unspecified.

showed a similar non-rhyme preference at a much higher frequency (27% vs. 20% for rhymes), although a battery of statistical tests generally failed to find significant differences in rhymes vs. non-rhymes.[11] As the asymmetric relative preference that /-s/ shows for non-rhymes, in contrast to all other plural suffixes, is core to the dual-route analysis that /-s/ is the only regular plural class in German (Clahsen, 1999b), Z&L's results appears to call this claim into question. Their findings are also presented and further discussed in Chapter 3.

[11]Grammatical gender *was* found to be significant, in keeping with earlier work (Köpcke, 1988), along with a number of other factors such as word-final phoneme. It's possible that grammatical gender and rhyme status could interact as factors, although this interaction does not appear to be incorporated into any of the analyses.

Chapter 3

German Speaker Data

Comparing the behavior of modern neural networks and human speakers on German plurals requires a detailed overview of speaker behavior. However, in spite of the research literature on this topic, at the time of this project no fine-grained data was available. In an appendix, Marcus et al. (1995) report average ratings for each plural category on each of the 24 nonce nouns tested; while helpful, this level of detail does not provide any indication of speaker variability, or permit analysis of speaker-item interaction. Zaretsky and Lange (2016) only report frequencies of plural category production across all 24 items, which reveal a different overall pattern from the earlier rating findings, but do not enable item-level analysis.[1] In addition, it is unclear whether these different outcomes might be at least partly due to differences in study design, as the later study used a production rather than rating task.

In light of these limitations, we opted to collect new German speaker data on the same stimuli to enable fine-grained comparison with neural networks, and compare production and rating to help clarify the role of potential task effects. The results show a strong quantitative preference for /-e/ and /-en/ over /-s/, and evidence consistent with both /-en/ and /-s/ as productive classes for these stimuli.

3.1 Study design

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. All materials can be found in Appendix A.

[1] We inquired after the data, but were informed by the authors that it was unavailable.

Case	SG-Masc	SG-Fem	SG-Neut	PL
Nominative	der	die	das	die
Accusative	den	die	das	die
Dative	dem	der	dem	den
Genitive	des	der	des	der

Table 3.1: German definite articles

3.1.1 Stimuli

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. For the second and third sections, the production and rating tasks, the twenty-four nonce words originally developed by Marcus et al. (1995) were presented. The stimuli comprised twelve nouns which rhymed with existing German words ("rhymes"), and twelve nouns which, while presumably phonotactically valid, did not rhyme with any existing German word and would likely be considered less phonologically natural ("non-rhymes"). Please see Table 2.3 for example stimuli.

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 presented in the introductory production task were selected for neuter gender, so they followed this pattern as well. This design decision had several motivations. The first was to exclude variation which might be attributable to grammatical gender — while Marcus et al. (1995) found no effect of grammatical gender on wug test behavior, Zaretsky and Lange (2016) reported a significant effect of feminine compared to non-feminine (masculine and neuter) grammatical gender. Another, related motivation was to provide a clear and consistent cue to noun number. While the article *die* can indicate either plural number or singular number with feminine gender, *das* can only indicate singular number; after a noun is presented with the article *das*, the following prompt starting with *die...* gives an unambiguous cue to plural number. Following this line of reasoning, neuter gender was selected over masculine gender because it provides a clearer signal to number across

different cases: while masculine gender singular articles might overlap in form with plural articles (i.e. *der* and *den*; c.f. Table 3.1), neuter singular articles never appear before plural nouns.[2]

3.2 Data collection

3.2.1 Presentation

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. Within each task, presentation order of items was randomized. All items within a task appeared on the same page, so participants could potentially see all items (e.g. by scrolling on the page) before responding to any prompt. Example images of stimuli presentation are included in Appendix A.

For the production task, participants saw the noun on its own, preceded by *Das*, e.g. *Das Bral*.[3] Above the response box, the text *Die...* appeared, to indicate that a plural form of the 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). While item order was randomized, within each item, the order of plural class presentation was fixed; see Appendix A for details. To check for participant attention, three items in the rating task contained an additional entry indicating one of the values of the scale (e.g. *Gut*; see Figure A.2 for an example). Participants were instructed to simply select the indicated value in these cases.

3.2.2 Participants

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

[2]This consideration was more important in an earlier iteration of this research plan, which sought to test morphological generalization of character-level neural language models as well, and thus required maximally unambiguous prompts to number in the form of preceding string sequences; see (Hahn and Baroni, 2019) for related exploration of language models' capacity to learn number cues.

[3]Note that this consistent presentation of neuter grammatical gender differs methodologically from both previous studies, which presented half of items as feminine gender and the other half as non-feminine, systematically varied across participants.

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

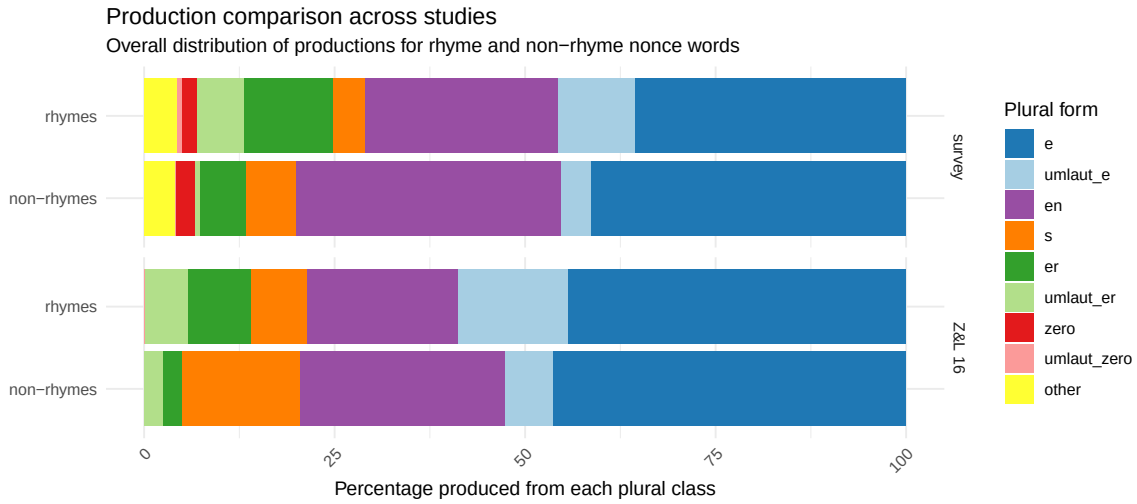

Figure 3.1: Comparison of plural type productions from the current study ("survey", top) and Zaretsky and Lange (2016, Table 2: "Z&L 16", bottom).

survey. We reimbursed participants at a minimum rate of £5/hour via the platform.

With the intent of releasing the data, we asked respondents only demographic questions that we thought might influence their judgment of plural forms. Respondents tended to skew young, with 80% reporting their age below 24. They also tended to speak a great deal of English: 92% reported at least a B1 level proficiency, and 62% said they interacted in English, or with English language content, 10 or more hours per week. These factors might be thought to encourage a bias toward /-s/ in our survey respondents, given its high frequency as the regular plural suffix in English. About half of respondents reported also speaking or learning another language than English or German.

After filtering out respondents who failed the attention check described above, data from 150 participants was available for analysis. The cleaned, anonymized survey data will be published online.

3.3 Results

3.3.1 Production

Figure 3.1 shows the overall distribution of different plural forms in the production task, comparing the data from the current survey with the findings reported by Zaretsky and Lange (2016). Rhymes and non-rhymes are presented separately for comparison.

The production data collected in our survey appears broadly consistent with the

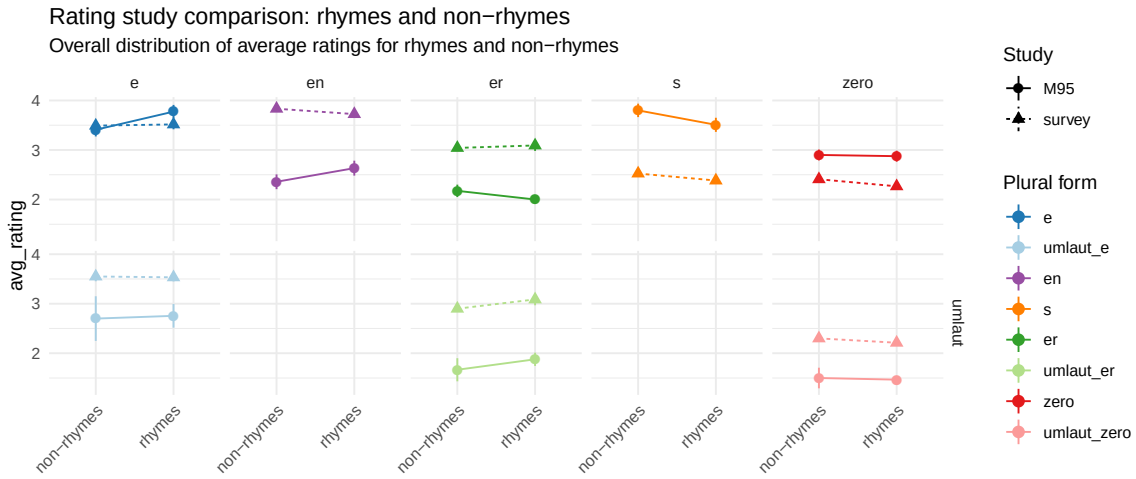

Figure 3.2: Comparison of average ratings studies from Marcus et al. (1995: M 95) and current study.

distribution observed by Zaretsky and Lange (2016) in their large-scale replication involving 585 participants. While some minor discrepancies are apparent — for example, it seems that none of Zaretsky and Lange's participants produced zero plurals or irregular forms outside of the established plural classes (labeled 'other' in the survey data) — the general distribution of plural category productions across the two studies is very similar; a Chi-squared test comparing the overall distributions of the two studies finds no reliable difference ($\chi^2 = 54, df = 48, p = 0.26$). In both studies, the plural form /-e/ is most favored for production (45% for Z&L, 38% in the current survey) with /-en/ a near second (23% Z&L, 30% here), while /-s/ is distinctly marginal (11% Z&L, 5% here). /-s/ is favored for non-rhymes (6%) over rhymes (4%), but, as in Z&L's results, /-en/ shows the same pattern with much higher frequency (35% vs. 25%). Even the suffix /-er/ (with or without umlaut), which is generally analyzed as a non-productive plural class (Dressler, 1999), is extended to non-rhymes at roughly the same rate as /-s/ (6.7% /-er/ vs. 6.4% /-s/)

3.3.2 Rating

Figure 3.2 shows the overall distribution of plural form ratings in the rating task, comparing the data from the current survey with the findings reported by Marcus et al. (1995) for "root" category stimuli. Rhymes and non-rhymes are presented separately for comparison.

The survey rating data show clear differences to the original Marcus et al. findings.

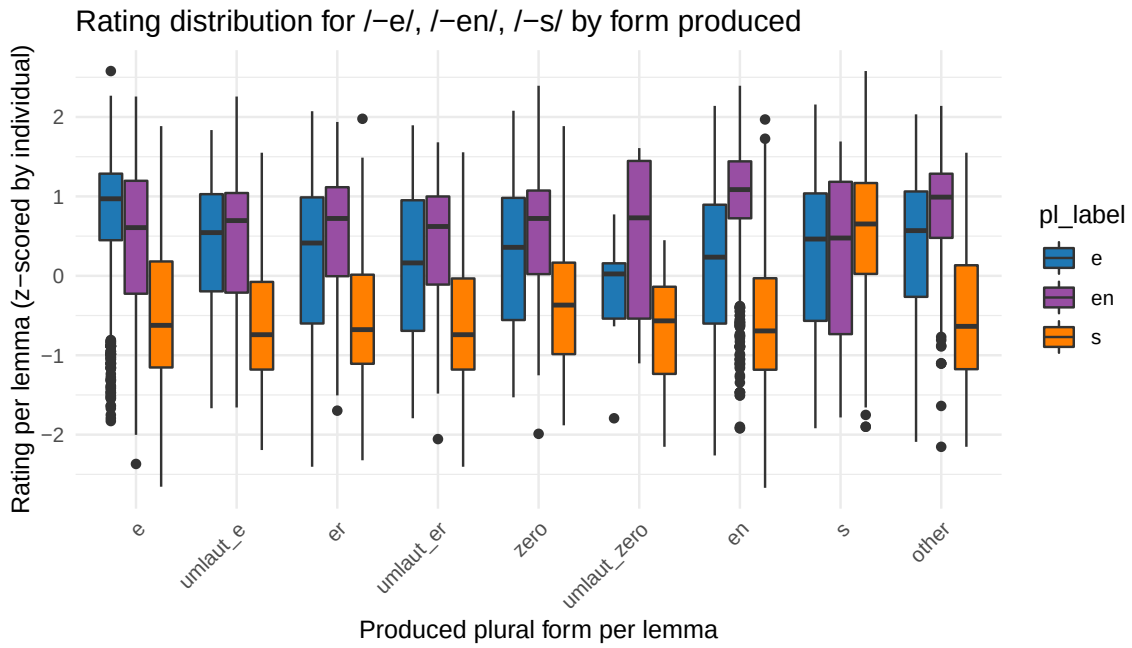

Figure 3.3: Overview of ratings for selected plural forms, grouped by the plural form originally produced on the speaker's first encounter with the nonce word (lemma).

The highest-rated plural form in the survey, /-en/ (average rating 3.8), was in the middle of the pack in the original ratings study (average 2.5) — and the highest-rated plural form in the original study, /-s/ (average rating 3.7), was among the three least preferred plural forms in our survey ratings (average 2.5).

3.3.3 Analysis

3.3.3.1 Comparing ratings and production

The production and rating data from the survey show an interesting asymmetry: although /-e/ is the preferred form in production with 38% compared to 30% for /-en/, /-e/ and /umlaut-e/ tie for second place in the ratings with an average of 3.5, compared to the 3.8 rating for /-en/ — indicating that participants appear to produce /-e/ more often, but rate /-en/ higher. Figure 3.3 illustrates how this asymmetry comes about. In the rating task, /-en/ is rated quite high across the board, including by speakers who produced /-e/ for the same word in the earlier production task; however, speakers who produced /-en/ tend to rate /-e/ much lower, possibly indicating a stronger comparative preference for /-en/ relative to /-e/. Another noteworthy aspect of this figure is the relatively weak comparative preference for /-s/: even speakers who originally produced /-s/ for a given word tend to rate /-e/ and /-en/ forms as roughly equally good. On

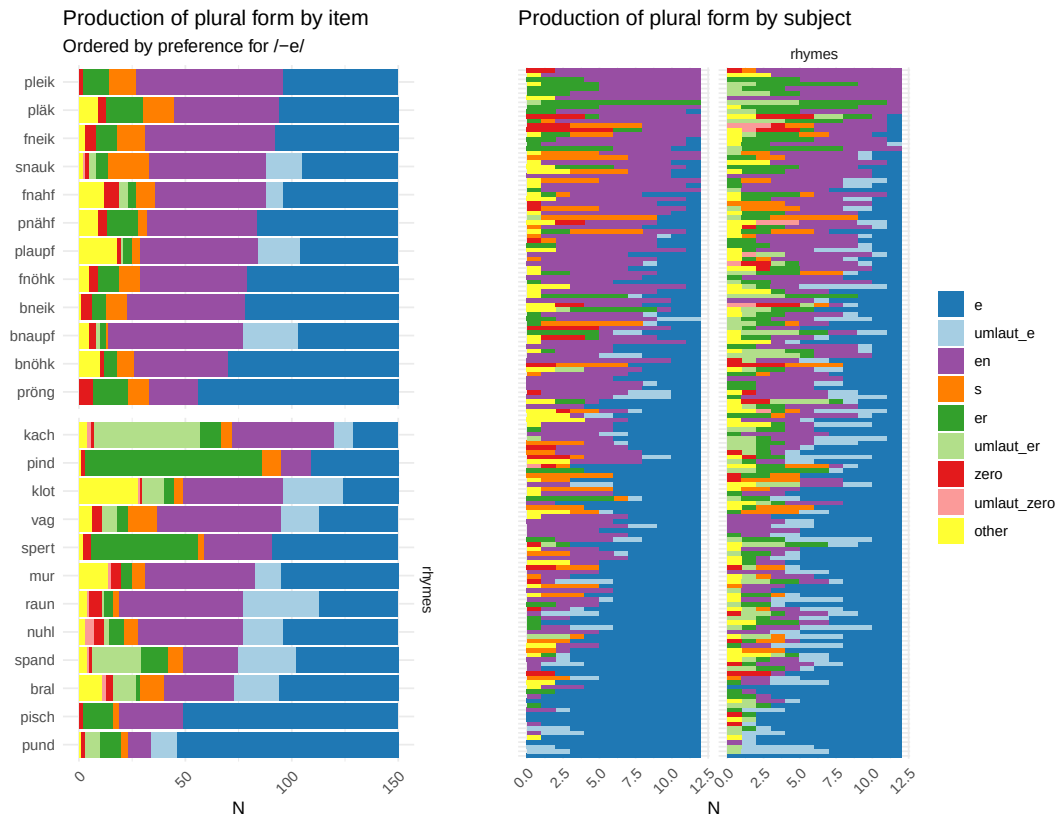

Figure 3.4: Item and subject variability in plural form production.

average, the top-rated plural form for a given word obtained 39% of productions for that word, with high variance ($\sigma = 13$).

On closer inspection, survey respondents appeared quite willing to consider plural forms that they did not produce as "good". In 32%, fully one-third, of all cases, participants rated one or more forms that they did *not* produce as *better* than the form they first used on encountering the same nonce word — and 97% of participants rated a non-produced form higher than a produced form at least once, so this trend was not driven by a small group of self-doubting speakers. Nonetheless, speakers still showed a broad alignment between rating and production, rating the plural forms they produced as 4.0 on average, higher than the global average of 3.0.

3.3.3.2 Item and subject variability in production

Figure 3.4 shows the distribution of plural form productions grouped by item (left) and by subject (right), ordered by preference for /-e/, the most commonly produced plural form. Both display a great deal of variability within individual items and speakers.

With respect to individual items, an overall trend is apparent: nonce words with

existing German rhymes, presumably more phonotactically 'typical' of the language, show greater diversity in produced forms, an effect which is discussed further in the following section. Nonetheless, the survey production data does not show much evidence for strong production preferences related to particular words, with a couple exceptions clearly influenced by frequent rhymes (i.e. *Pund – Punde* on analogy to *Hund – Hunde*, *Pind – Pinder* from *Kind – Kinder*, *Kach – Kächer* from *Dach – Dächer*; the last two may even reflect experimenter error, as those two rhyming forms appeared in the practice test with real plurals, possibly priming the survey respondents). Notably, even though /-e/ is produced 38% of the time overall, it doesn't predominate in any particular item. /-e/ shows the highest associations with *Pisch, Pund*, and *Pröng*, with roughly 66% production probability for each of these words, and around 50% for *Bnöhk*. For the rest of the words, preferences for /-e/ are reliably at or below 40%.

Speaker preferences show even higher variability, ranging smoothly from 100% use of /-e/ to 0%, with about almost a third of speakers using /-e/ on more than half of the items, and a quarter of speakers using /-e/ on less than a quarter of the items. There is an apparent trade-off between /-e/ and /-en/, as speakers who produce /-e/ less use /-en/ more and vice versa. While the general trend for greater diversity on rhyme nonce words is visible within individual speakers as well, no consistent patterns of application or clusters of speakers are immediately apparent.

3.3.3.3 Statistical analysis: Subject and item factors

The effects of word familiarity (i.e. having existing German rhymes) and subject demographic factors (English proficiency level, regular exposure to English, age) on the survey data was evaluated using mixed-effect models fit with the *lme4* package in R (Bates et al., 2015; R Core Team, 2019). As umlaut production is statistically confounded with word familiarity in the stimuli (only words with a back vowel can undergo umlaut, and 9 of the 12 rhymes have back vowels compared to 4 of the 12 non-rhymes), we modeled production and rating of plural suffixes, including umlaut and non-umlaut forms.[5] For each plural suffix, we modeled production likelihood with a binomial generalized linear mixed-effects model (GLMM) using the one item factor and three subject factors as fixed effects, with random intercepts specified for subject and item. Similarly, ratings for each plural suffix were analyzed using a linear mixed-effects model (LMM) with the same effect structure. Ratings were represented

[5]Follow-up models which separately analyzed umlaut and non-umlaut plural forms showed similar results to the ones reported here for suffixes.

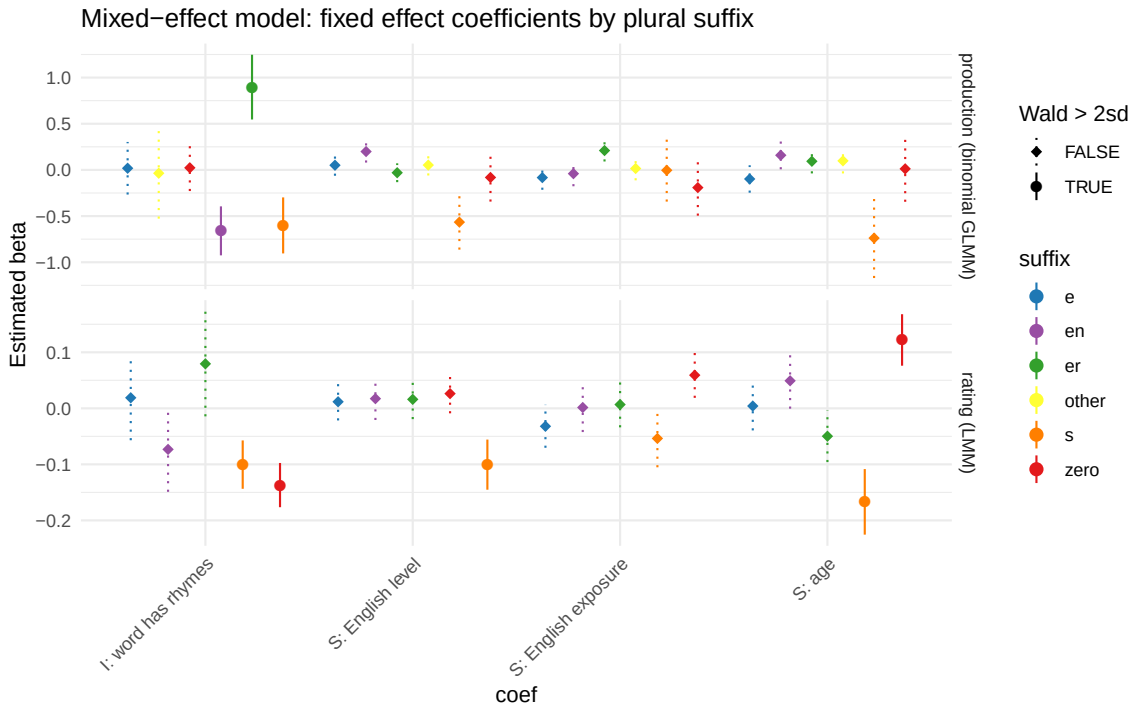

Figure 3.5: Estimated effect of item (I) and subject (S) factors on production (top) and rating (bottom) for each plural form.

as z-scores scaled for each participant.

Model results are presented in Figure 3.5. Coefficients with a Wald statistic (z for the production GLMM, t for the rating LMM) greater than 2 standard deviations from the mean, a rough proxy for statistical reliability, are indicated with round points and solid lines.

For the production task, while subject demographic factors do not have any reliable impact, word familiarity has effects on three suffixes: subjects are more likely to produce /-er/, and less likely to produce /-s/ or /-en/, on words with German rhymes compared to non-rhymes. The rating task shows a similar pattern in the direction of effects, but only /-s/ and zero-suffix forms are reliably rated higher for non-rhyme words than rhyme words. These suffixes also show effects of subject demographic factors on their ratings: older respondents tend to rate zero-suffix forms higher, and /-s/ forms lower. Perhaps surprisingly, higher English proficiency is also associated with lower ratings for /-s/ forms. Furthermore, although the effect is not statistically reliable, the same trend appears in the production task, and in the relation of regular English exposure to /-s/ rating — it appears that higher English experience steers participants away from /-s/ plurals in German, perhaps influenced by an inductive bias toward mutual

exclusivity (Markman and Wachtel, 1988; Gandhi and Lake, 2019).

3.4 Discussion

The survey data presented here support a few intermediate observations:

Preference for /-s/ may be diminishing over time. While Köpcke (1988) found that subjects produced /-s/ for 14% of non-feminine monosyllabic nouns, and M95 found that /-s/ was rated highest on average for 9 of the 24 nonce nouns (37%), Z&L's research (conducted 2011-2013) identified an overall production rate of 11% for /-s/, and participants in the current survey produced /-s/ in a mere 5% of cases, and never rated /-s/ highest for any individual item. This may reflect a historical shift: Elsen (2002) has noted that some loanwords in German which originally took /-s/ have shifted to /-e/ or /-en/ (e.g. *Pizzas* → *Pizzen, Kiosks* → *Kioske*). The impact of second languages, especially English, has long been speculated to affect /-s/ use in German (e.g. Pulvermüller, 1999; Stemberger, 1999). Our analysis of demographic factors suggests this effect may go in the opposite direction than usually assumed: younger speakers and speakers with higher English proficiency actually appear *less* likely to produce, or rate highly, plural forms with /-s/.

Any evidence for /-s/ default also points toward /-en/ default. Although /-s/ shows the predicted effect of word familiarity on production and rating, with a stronger association to non-rhyme nonce words in both tasks, /-en/ shows the same direction of effect, but with far higher production probability and average rating. This production finding is also consistent with Z&L's results. Notably, even speakers who *produced* an /-s/ plural for a given word typically rate /-e/ and /-en/ as roughly equally good for that same word. The non-rhyme preference of /-s/ on these stimuli is considered crucial evidence for its singular status as the regular default (Clahsen, 1999a). From our survey data, it appears that the case for /-en/ on these stimuli is just as strong, if not stronger due to its greater overall frequency.[6]

Production and rating show broadly consistent but distinct results. While participants give higher ratings on average to the same plural forms they produce on a given nonce word, they also give higher ratings to other plural forms in one-third of cases. One unexpected outcome of the cross-task differences is an asymmetry in pre-

[6]Marcus et al. (1995, 232) justify treating /-en/ as irregular in their study as "it clearly is not rule-generated in the case of monosyllabic nouns"; regardless of whether it's rule-driven behavior, our survey results clearly indicate /-en/ is productively extended to such nouns.

ferred forms: while /-e/ is produced for 38% of nonce words overall (45% including forms with umlaut, compared to 30% for /-en/), /-en/ is the most highly rated plural form by a relatively wide margin (3.8, compared to 3.5 for /-e/). This result highlights the potential for different methodologies to support divergent conclusions.

Overall, the survey data presents a complex and highly variable picture of German speaker preferences for plural forms. While discrepancies between the production and rating data for the same survey participants point toward task effects which might contribute to the differences between the Marcus et al. (1995) rating data and the Zaretsky and Lange (2016) production data, the original claim of a preference for /-s/ finds even less support in the current survey across tasks. /-e/ and /-en/, the most frequent plural classes observed in German nouns generally, retain their predominance in both production and rating for this wug test — but as Figure 3.4 illustrates, variation within items and speakers appears to be the rule rather than the exception. This diverse distribution presents a challenge for modelling: will neural networks trained on German nouns show a similar range of behavior? The next chapter explores this question.

Chapter 4

Encoder-Decoder Model

4.1 Encoder-Decoders for morphology

The task of mapping a singular-number input string to a plural-number output string is a special case of the more general problem of morphological *transduction*, or mapping from some input (traditionally the *lemma*, or citation form of a word) to one or more outputs with some specified inflection (e.g. Clark, 2002; Dreyer and Eisner, 2011; Durrett and DeNero, 2013; Cotterell et al., 2016). Currently, neural models with encoder-decoder architectures achieve state-of-the-art performance on the general task of morphological transduction (Kann and Schütze, 2016; Cotterell et al., 2018). Previous models of the German plural system have been dismissed for evaluating a simpler classification task as opposed to transduction proper (Hahn and Nakisa, 2000; Pinker and Ullman, 2002); modern neural transduction models should resolve this criticism.

Encoder-decoder (ED) models comprise two Recurrent Neural Networks (RNNs) combined to perform sequence-to-sequence transduction (Sutskever et al., 2014). The encoder RNN processes each symbol in the input string sequentially, yielding at each time step a fixed-length vector representation (a *hidden state*) conditioned on the current input symbol and the representation from the previous sequence of inputs, with the final time step producing a vector encoding the entire sequence. The decoder RNN performs a similar sequential operation, conditioning on the encoded representation and any previous output sequence at each time step to produce the next output symbol. Typically, the decoding process is additionally coupled to the input via an *attention* mechanism (Bahdanau et al., 2015): rather than maintaining a fixed encoding of the input, the decoder can learn a vector of attention weights over the sequence of encoder hidden states, so that the representations of certain inputs can exert greater influence

over particular decoder steps.

In recent work, Kirov and Cotterell (2018, henceforth K&C) argue that modern ED models can successfully approximate language learners for cognitive modeling purposes. K&C offer two key claims to support this view: 1) ED models' success on novel inputs implies they induce correct *generalizations* about the data, and 2) their *errors* are human-like, showing they mimic speaker behavior. The authors turn to the long-studied English past tense (e.g. Rumelhart and McClelland, 1986; Pinker and Prince, 1988, inter alia) to make their case, training an ED model to map phoneme-level representations of English verb stems to their past tense inflected forms (using data from CELEX: Baayen et al., 1993). On a held-out test set of existing English verbs, the model appears to support both claims: it *generalizes* the regular past tense suffix /-d/ (with appropriate phonological variation in voicing), achieving near-ceiling accuracy on unseen regular verbs, and its *errors* on unseen irregular verbs all result from overregularization[1], a process consistent with human behaviour (Albright and Hayes, 2003). K&C additionally evaluate the model through wug-testing on a set of 74 English nonce verbs, and find that speakers' production probabilities of particular regular and irregular forms were more correlated with ED model scores than the rule-based learner proposed by Albright and Hayes (2003).

In follow-up work, Corkery et al. (2019) call K&C's wug-test results into question. They find that the model is highly sensitive to initialization conditions, as simulations with different random seeds on the same architecture yield considerable variation with respect to correlation to speaker production probabilities — with most results showing a lower correlation than reported by K&C, weakening the *generalization* claim. They also find that individual simulations tend to rank implausible nonce forms highly (i.e. the model's *errors* may also be less human-like than claimed), although these distinctions are smoothed out when the model results are aggregated across simulations; nonetheless, the original rule-based learner shows a higher correlation to speaker production probabilities than the aggregated results of the ED models, raising uncertainty as to whether ED models in fact represent a superior choice for cognitive modeling.

While the work of K&C and Corkery et al. (2019) usefully updates the past tense debate in light of the current generation of more powerful computational models, it also inherits one of the main limitations of that debate — namely, the focus on the English past tense, an inflectional system with a high degree of simplicity that is far

[1] Although the authors note four errors on regular verbs involving vowel substitution within the stem, a phenomenon much less likely to be produced by speakers.

from representative across languages (Seidenberg and Plaut, 2014). As 96% of the verbs in K&C's subset of CELEX take the regular past tense /-d/ (modulo voicing), it would be quite an indictment of the model should it *fail* to generalize that class. The lopsided distribution of classes reduces the impact of the generalization claim for this phenomenon, leaving the error claim to carry more evidential weight in the ED-models-as-learners argument. However, it is unclear how best to evaluate this claim. Corkery et al. (2019) note the tendency of their ED models to place the bulk of probability mass on the top item in the beam (generally the regular form), which means that small shifts in beam probability for irregular forms can have large impacts on ranking and correlation measures. In addition, it's unclear how to evaluate model outputs that are unattested in speaker data, as seen in Table 1 of the same paper. How much less 'human-like' are unobserved forms such as *nold–nelt*, *nold–neelded*, or *murn–murn*, in comparison to observed forms such as *nold–neld* and *murn–murnt*? The wug tests of Albright and Hayes (2003) were designed to elicit novel inflected forms from speakers; when ED models produce unattested forms, what does it tell us about their capacities as learners?

As Marcus et al. (1995) observed, the German plural system shows a much broader distribution over classes (c.f. Table 2.1); regardless of whether any of the plural classes count as 'regular', any frequency-regularity confound simply cannot hold with no class holding the majority. For this reason, German plurals present a more suitable inflectional system to assess K&C's generalization claim, as the task of correct generalization will be more challenging. The wider distribution of plural classes could also make errors more informative, as each input has a broader range of plausible outputs.

4.2 Methodology

4.2.1 Model

Implementation: Following K&C and Corkery et al. (2019), our model is implemented using OpenNMT (Klein et al., 2018) with their reported hyperparameters (originally following Kann and Schütze (2016)): 2 LSTM encoder layers[2] and 2 LSTM decoder layers, 300-dimensional character embeddings in the encoder, and 100-dimensional

[2]K&C used bi-directional encoder layers, but due to experimenter error, the model described here is identical in all respects *except* that the encoder layers are uni-directional. Follow-up investigation with 25 simulations of the correctly-specified biLSTM model found that its performance did not differ substantially from the results reported here, with 87.9% accuracy on the test set compared to 87.3% for the described model; accuracy was also roughly the same within each plural class.

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.

Random simulations: As 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. In keeping with the rough heuristic that an individual simulation, in a cognitive modeling scenario, might be considered roughly analogous to an individual speaker, we trained 150 unique random seeds, so that the order of magnitude of productions would be comparable to that of the 150 survey participants.

4.2.2 Data

The ED neural model for morphological transduction takes character-level representations of German nouns in their singular form as inputs, and learns to produce the noun's inflected plural form as the target output. In contrast to English, the phonological-orthographic mapping is quite straightforward in German, which enables us to train on character sequences directly. In that respect, this work contrasts with related work on the English past tense (Rumelhart and McClelland, 1986; Corkery et al., 2019; Kirov and Cotterell, 2018) and on German plurals (Nakisa and Hahn, 1996; Hahn and Nakisa, 2000), all of which used sequences of phonemes or phonetic feature representations.

Corpus: We trained all models on the Unimorph German data set[3] (Kirov et al., 2016; Sylak-Glassman et al., 2015), which provides the singular and plural forms of 11,243 nouns. Only nominative case forms of each noun and number combination were used. Note that the training data included only bare nouns — unlike the survey participants, the model received no cues to grammatical gender (or case, or number) in the form of a preceding determiner.

Splits: To assess generalization on existing German nouns, every tenth noun was extracted into a held-out test set. Marcus et al. (1995) provided a list of German nouns which rhymed with the 'rhyming' wug stimuli (e.g. *Tisch, Fisch, Wisch* all rhyme with the nonce form *Pisch*); to ensure that the model was exposed to enough forms to generalize correctly, we checked that none of these frequent rhyming nouns appeared only in the test set and not in the training set.[4] As hyperparameters and other implementa-

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

[4] In checking this, we counted compound words in which the rhyming noun formed a head, e.g. *Arbeitstisch, Stammtisch* for *Tisch*, as instances of that noun; otherwise we did not perform any reduction of the corpus to control for frequency of different compound heads, in contrast to some other research work (Hahn and Nakisa, 2000).

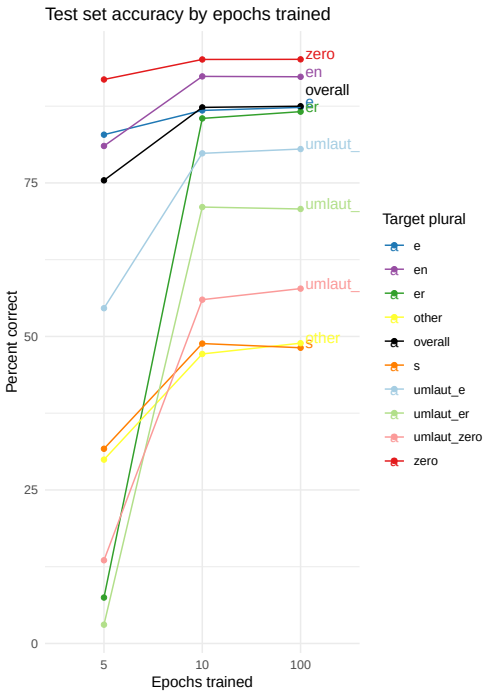

Figure 4.1: Average test set accuracy by plural class at 5, 10, and 100 epochs.

Pl class	Train	N	Test	N
/-(e)n/	100 (0)	3731	92.3 (0.6)	462
/-e/	100 (0)	2791	86.8 (1.3)	301
+uml	100 (0)	715	79.8 (3.6)	66
/-ø/	100 (0)	1853	95.1 (1.1)	204
+uml	100 (0)	92	56.0 (9.3)	11
/-er/	100 (0)	74	85.6 (5.3)	16
+uml	100 (0)	207	71.1 (5.4)	32
/-s/	99.8 (0)	413	48.9 (4.7)	35
other	99.5 (0.1)	214	47.2 (4.7)	26
overall	100 (0.2)	10090	87.3 (0.5)	1153

Table 4.1: Train and test set accuracy in percent (std. dev.) by plural class, averaged over 150 simulations.

tion details were specified based on the literature, no model tuning was anticipated and hence no separate development set was extracted.

Epochs: K&C trained their model for 100 epochs to ensure all irregular forms were memorized;[5] Corkery et al. (2019) followed this procedure, but reported the highest correlation between model predictions and wug-test productions after only 10 epochs, even though 60-80 epochs was required for the small set of English irregular verbs to reach ceiling accuracy. In the current experiment, the model appeared to memorize the training data within 10 epochs, reaching an average accuracy of 99.98% on the training data and 87.3% on the held-out test data. To verify that 10 epochs was not underfitting, we trained 15 additional simulations until 100 epochs, which achieved an average accuracy of 87.5% on the test set. We conclude that 10 epochs should suffice for evaluation, and all reported results are for models trained to 10 epochs. See Figure 4.1 for accuracy by plural class on the test set at 5, 10, and 100 epochs of training.

[5] Although memorizing the training data would typically be considered overfitting and thus avoided in machine learning research, for cognitive modeling we want the training data to be memorized, as we expect adult human speakers to have perfectly mastered all forms in their vocabulary; in this context, performance below 98% or so would be underfitting.

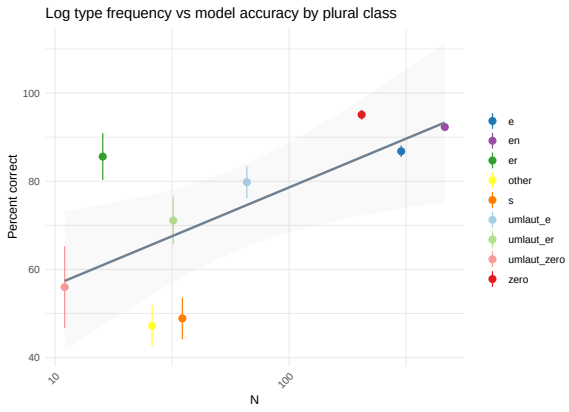

Figure 4.2: Log frequency and accuracy by plural class.

Pl suffix	Baseline	Model
/-(e)n/	75.5	92.3 (0.6)
/-e/	75.5	85.5 (1.2)
/-ø/	56.7	93.1 (1.1)
/-er/	0	75.9 (4.0)
/-s/	34.3	48.9 (4.7)
other	0	47.2 (4.7)
overall	65.9	87.3 (0.5)

Table 4.2: Average model accuracy (std. dev.) compared to frequency baseline by plural suffix.

4.3 Results

Table 4.1 reports training and test set accuracy by plural class. Reported accuracy metrics show recall relative to the *target* plural class, e.g. 95% accuracy for the zero plural class /-ø/ indicates that, of the 204 forms in the test set taking a zero plural, the model produced the correct zero-inflected form for 95% of them.

A detailed investigation of ED model outputs in comparison to speaker data will be provided in the following chapter. Here, we present an overview of the model's outputs, highlighting a few selected outcomes of note.

4.3.1 Test data

Test data performance quantifies the model's capacity to correctly predict the plural form for unseen German nouns, thereby providing an evaluative measure for K&C's *generalization* claim that ED models induce correct biases. The average test set accuracy overall is 87.3%, a fair bit lower than the 95% reported by K&C for the English past tense, although higher than the 83% accuracy Hahn and Nakisa (2000) achieved with a feed-forward neural net on the simpler task of plural classification. Within the plural classes, test set accuracy ranges from 95% (for zero-suffix forms) to 47-48% (/-s/ and other forms); all are below the 99% test set accuracy of K&C's English regular verbs, but all classes are also comfortably above the 29% test set accuracy for English irregular verbs.

Baseline comparison: We evaluated the model's performance against a baseline

strategy of picking the most frequent plural class observed in the training data given the final character of the input word.[6] As this baseline is unable to distinguish an input's capacity to undergo umlaut, we computed accuracy over suffixes only. Table 4.2 shows the results. Given that the strategy of always picking the most frequent plural class overall (/-en/) yields 40% accuracy on the test set, the frequency by last character baseline represents a substantial improvement, yet one that the model readily beats.

/-s/ and other: One readily apparent result of the test data presented in Table 4.1 is that the model performs worst on the plural classes /-s/ and the pseudoclass 'other'. In the case of 'other', this is not particularly surprising, as it refers not to a class, but to the long tail of idiosyncratic inflections that do not fall into one of the existing plural classes. The plural class /-s/ is a more interesting case: the model's failure to generalize /-s/ correctly is prima facie consistent with the Marcus et al. (1995) analysis of /-s/ corresponding to the 'elsewhere' condition, and thus resistant to data-driven generalization.

One potential explanation is that /-s/ and 'other', at 3% and 2.2% of the corpus respectively, are simply too infrequent to support effective generalization. However, Figure 4.2, which visualizes a linear model fit to log frequency and accuracy by plural class, shows that performance on these two categories is still lower than expected given type frequency in the corpus. Again, this finding is unsurprising for the catchall 'other', and more noteworthy with respect to /-s/.

Headedness: A well-known aspect of German which could hypothetically aid a model's capacity to generalize is the variety of compositional mechanisms which form complex nouns. Two phenomena of particular interest are *derivational suffixing* and *compounding*. Derived nouns, such as *Reduzierung* from *reduzieren* + *-ung* (English: *reduction* from *reduce* + *-tion*), inherit the plural class of their derivational suffix, i.e. /-en/ for *-ung*; similarly, compound nouns inherit the grammatical gender and plural class of their head noun, the rightmost noun in the compound. The UniMorph dataset features a large number of compounds and derived nouns — does the ED model learn to exploit this phenomenon and extend the plural class of nouns observed in training to compounds with shared heads in the test set?

Table 4.3 gives some examples of compound and derived nouns found in the test set which share a derivational head with one or more words in the training set. Derivational suffixes appear to be learned quite well: of the suffixes tested by Köpcke (1988),

[6] While unsophisticated, note that the strategy of predicting the most frequent plural class overall achieves 95% accuracy on the English past tense data.

Pl class	Head	Train	Test	Acc(%)
/-s/	*Chef*	*Fraktionschef, Staatschef, Vorstandschef*	*Parteichef*	89.3
/uml+e/	*Saft*	*Fruchtsaft, Apfelsaft* and 9 more	*Waldmeistersaft*	78.7
	Hut	*Fingerhut, Erzherzogshut*	*Sonnenhut*	51.3
/-er/	*Kind*	*Einzelkind, Kleinkind* and 2 more	*Stiefkind*	92
	Rind	*Rind*; c.f. *Grind-e*	*Auerrind*	27.3
/-e/	*Hydrat*	*Hydrat, Dihydrat* and 5 more	*Tetrahydrat*	100
/uml+ø/	*Laden*	*Fensterladen, Gemüseladen*; c.f. *Fladen-ø*	*Buchladen*	70.7
			Lebensmittelladen	52.7
			Bioladen	19.3
/-en/	*-krat*	*Autokrat, Demokrat* and 3 more	*Bürokrat*	63.3
/-ø/	*-lein*	*Büchlein, Welplein* and 6 more; c.f. *Bein-e, Schein-e*	*Kindlein*	100
			Knäblein	94.0
			Fräulein	82.0

Table 4.3: Examples of compound (top) and derived (bottom) nouns in training and test data.

-ling, -ung, -chen reach over 99% accuracy in the test set; as shown in the table, *-lein* is the most variable, with a mean test set accuracy of 94%.[7] Spot-checking other common suffixes (*-heit, -keit*) reveals similarly high performance, although less common suffixes like *-krat* do not necessarily fare so well. In lieu of more detailed analysis of compound nouns, the table shows a wide range of accuracy on test data with the same compound head — while *Hydrat* generalizes perfectly to *Tetrahydrat*, *Laden* shows much greater variation: 70% of individual simulations produce the correct plural for *Buchladen*, but only 20% do so for *Bioladen*. Further investigation suggests that the presence of other words with the same final string but differing heads with differing plural classes, e.g. *Gaden-ø, Fladen-ø* for *Laden* or *Rat-uml+e, Karat-e* and 50 more for *-krat*, unsurprisingly lead toward greater instability of model predictions. It seems that the ED model has not learned to represent headedness, relying instead on the informative but noisy proxy of final string sequence.

One could imagine compound noun segmentation being learnable from orthographic sequences alone; however, the model is also at a disadvantage in this training regime, as it lacks information such as the noun's determiner, which could provide cues to heads of compounds and derived nouns, inter alia. K& C report higher performance with a model trained to jointly model multiple inflected forms, for example mapping

[7] The other suffix, *-schaft*, happened not to appear in the test set.

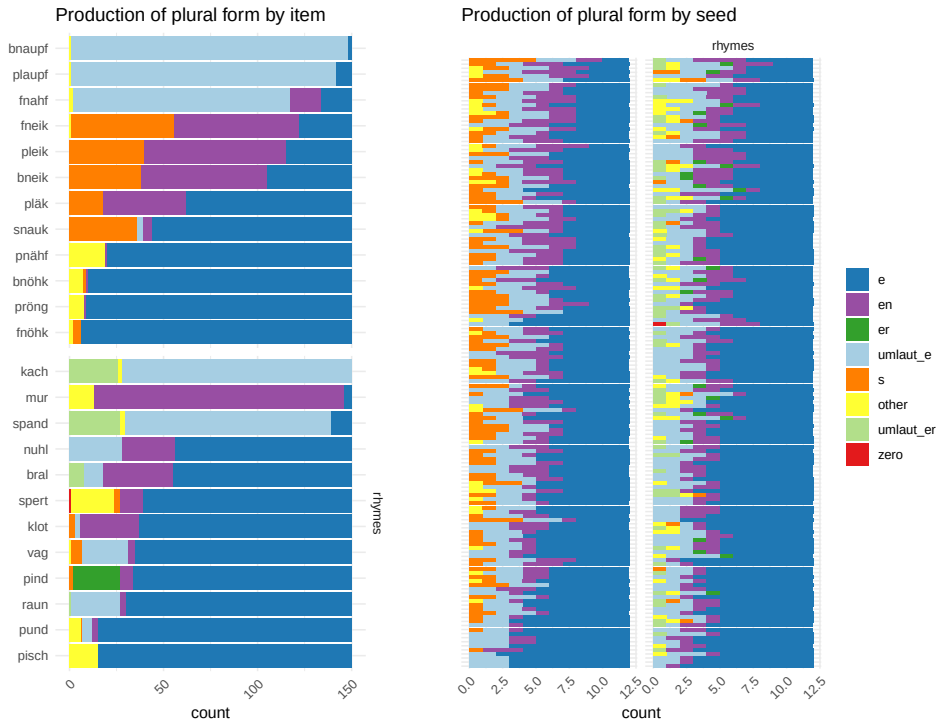

Figure 4.3: Variation in plural form production by item and random seed.

verbs not only from present to past tense, but to other inflections for tense, person, and number. In German plurals, one could train a mapping from singular to plural but also from plural to singular, nominative to dative case, and so on. This approach may also improve the model's ability to represent headedness, as evidence from various inflectional paradigms could support converging distributions for nouns sharing a derivational head.

4.3.2 Wug data

Figure 4.3 shows the distribution of plural forms predicted for the Marcus et al. (1995) wug test data by item and by random seed, comparable to Figure 3.4 for survey participants. An interesting contrast to the survey data is readily apparent: there is much more consistency across individual model simulations than across individual speakers — and, somewhat paradoxically, these more consistent item-based preferences lead to greater variability across items. The next chapter will go into more detailed comparison of the wug test results from survey participants compared to the ED model.

Chapter 5

Comparison and Analysis

This section compares the results of wug-testing for humans (via the survey data) and the Encoder-Decoder (ED) model, focusing on production probabilities. In general, the model appears to prefer the plural class /-e/ (with and without umlaut) very heavily relative to speakers, who produce a wide range of plural classes. /-en/ in particular is produced almost as often as /-e/ by speakers, but very rarely by the model. Production probabilities appear to correlate reliably between speakers and the model for /-s/ and /-e/ classes.

5.1 Production probabilities

Figure 5.1 shows the overall distribution of produced plural forms in the survey data compared to the ED model. One clear point of comparison is that the model prefers the /-e/ plural suffix to a far greater extent than speakers do: 54% of model predictions overall are for /-e/ without an umlaut (compared to 38% of speaker productions), and 20% of model predictions are for /umlaut-e/, far more than the 7% produced in the survey. In general, the ED model predicts a suffix other than /-e/ only 25% of the time, while 55% of speaker productions are forms other than /-e/.

As the model over-predicts /-e/, it also under-predicts /-en/: where speakers produce /-en/ plurals 30% of the time overall (for 35% of non-rhymes and 25% of rhymes), the model predicts /-en/ forms in 15% of cases regardless of word familiarity. This contrasts to the case of /-s/: speakers produce /-s/ plurals reliably more for non-rhymes (6%) than for rhymes (4%), and the model appears to pick up on this, predicting /-s/ in 11% of non-rhyme cases vs. 1% of rhymes, although see the next section for more detailed discussion.

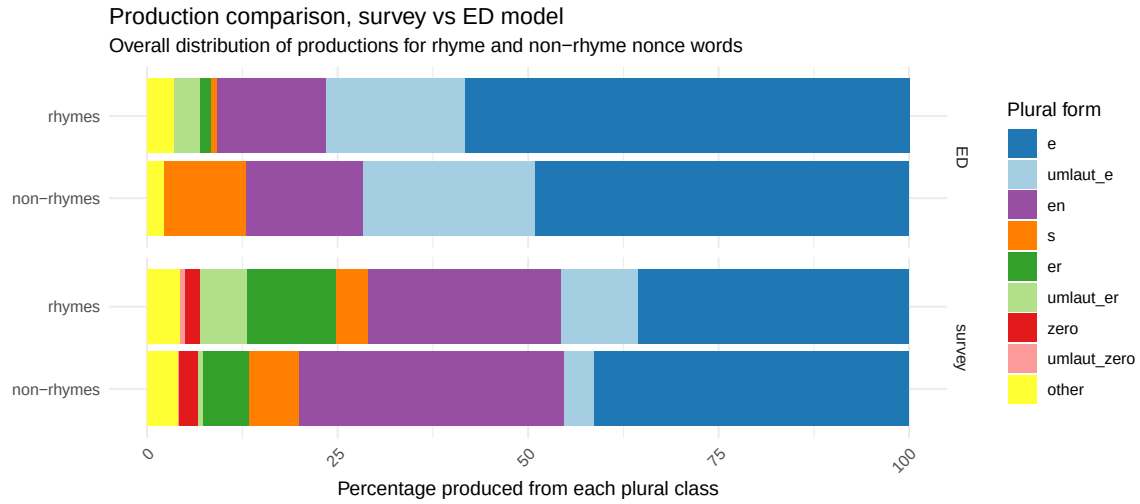

Figure 5.1: Comparison of plural type productions from the current study ("survey", bottom) and ED model (top).

In spite of these quantitative discrepancies, a Spearman rank correlation finds broad agreement between the production probabilities aggregated by item from the model and from survey participants. Table 5.1 shows Spearman's ρ for each plural class. Production probabilities for /-e/ and /-s/ appear to correlate reliably, even after certain /-s/-attracting items are excluded (see next section). The model also produces /-er/ (with or without umlaut) only for rhymes, which is consistent with the behavior of survey respondents, who produced /-er/ much more for rhymes. In general, model production probabilities across items correlate quite well with speaker production probabilities, with $\rho = .62$ overall (or .59 excluding certain items, see next section). Certain classes do not appear to correlate reliably, though, including /umlaut-e/, zero plural, and most notably /-en/.

Figure 5.2 presents an alternative visualization of ED model vs. survey participant productions for the plural classes /-e/, /-en/, /-er/, and /-s/, emphasizing the greater diversity of outputs from speakers. The x axis, with reversed scale, indicates W, the number of unique words for which a given plural class was produced. The y axis indicates N, the lower-bound number of individuals (participants for the survey, randomly-initialized simulations for the model) producing that class for W unique words. The overall chart roughly approximates a cumulative density function for agreement across individuals regarding particular words for each plural class. For example, the solid dark blue line in the top panel shows that there are (W) 20 distinct words for which at least (N) 41 speakers produced the plural class /-e/, while the dashed dark blue line

shows that only (N) 11 ED simulations applied /-e/ to that many words. Further to the right in the same panel, the dashed dark blue line rises above the solid dark blue line, indicating that a greater number of model simulations are applying the plural class /-e/ to a smaller number of words relative to the speakers. At W = 5, there are 5 distinct words for which 135 or more ED simulations predict /-e/, whereas speaker preferences are less concentrated: the most agreement in speaker /-e/ production for some set of 5 words is 72. The bottom panel shows /-er/ and /-s/. Note that at least one speaker produces /-s/ and /-er/ forms for each of the 24 nonce nouns, while the model produces /-s/ at all for only 12 distinct words; on the other hand, for a certain subset of three words (on which more below), 38 simulations yield /-s/ (dashed orange line), while at most 14 speakers overlap in producing /-s/ for three words (solid orange line). The main point illustrated by this plot is the extent to which speakers generalize plural classes across items, and fail to concentrate upon specific items to the extent the models do.

5.1.1 Item associations

/-eik/: Looking at individual items (Figure 3.4 for speakers, Figure 4.3 for the model), it's apparent that the model's word-familiarity effects on /-s/ production are highly driven by a few key items — *Fneik, Pleik, Bneik* and *Snauk* with 37%-24% /-s/ predictions, and to a lesser extent *Pläk* (12%), all notably words that end in *k*. /-en/ productions on non-rhymes also appear largely driven by the three-*eik* items.

A closer look at the training data resolves the mystery. *-eik* words do, in fact, rhyme with an existing German word in the training set: *Generalstreik*, which takes the /-s/ plural. Of the other 16 words in the training set which end in *-ik*, the loanword *Schaschlik* takes /-s/ , the loanword *Zaddik* takes the irregular form *Zaddikim* [1], and 14 others take /-en/. While the goal of the non-rhyming stimuli was to elicit a generalization response to atypical phonological inputs, it appears that the model's tendency to prefer /-s/ for non-rhyming stimuli — a behavior that would align with the generalization patterns of speakers[2] — is actually driven by several hidden rhymes. Table 5.1 further reports the correlation between model and speaker production probabilities with these rhyming non-rhymes excluded. /-s/ production continues to show a relatively high correlation ($\rho = .45$), but the correlation with /-en/ vanishes, dropping from

[1] Although some online searching reveals that this plural appears to alternate with the regularized form *Zaddiken*, e.g. `https://www.spin.de/forum/645/-/380e`.

[2] On re-analysis of speaker behavior with the /-eik/ words excluded, the tendency for /-s/ and /-en/ to prefer non-rhymes in production remains, although the statistical significance of the effect is diminished.

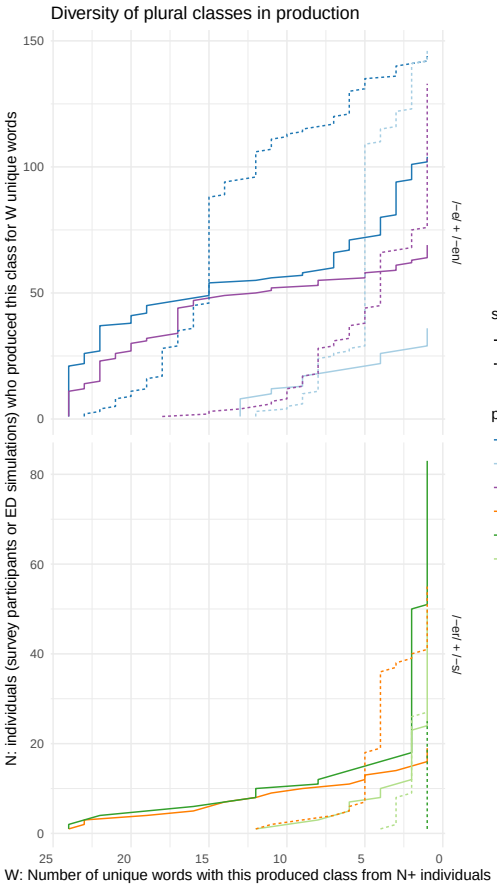

Figure 5.2: Comparing diversity in plural class production over items between survey participants (solid line) and ED model (dashed line). Note reversed x scale.

Pl class	ρ	ρ -'eik'
/-s/	.6**	.45*
/uml-er/	.55	.55
/-e/	.47*	.54*
/-er/	.34	.37
/-(e)n/	.24	-.02
/-ø/	.06	.09
/uml-e/	.04	.04
other	-.09	-.17
overall	.62**	.59**

Table 5.1: Spearman's ρ between ED and speaker production probabilities by plural class. Stars indicate p-value below .05 and .01. Right side shows recalculated values excluding *Fneik, Pleik, Bneik*.

.24 to -.02.

Mur: Another case of strong specific item associations from the ED is *Mur*, which 88% of simulations label /-en/, an anomaly for a plural class generally predicted for only 15% of the words overall (compare to speakers: 35% produced *Muren*, relative to 30% /-en/ production overall, and 37% of speakers produced *Mure*). Digging into the training data reveals a likely source — the very similar word *Mure* 'mudslide', which takes /-en/. This, combined with the statistical tendency of other words ending in *-ur* (63% of which take /-en/), likely accounts for the model's propensity to generalize /-en/ for this word in particular, while /-e/ remains the favorite everywhere else.

5.2 Assessing model scores

As discussed in Chapter 3, the relationship between speaker production (which presumably selects the best form) and rating data (which presumably ranks the relative acceptability of several forms) is somewhat unclear, with different plural classes favored in the production and rating task. Similarly, the relationship between ED model productions and the scores assigned to its lower-ranked forms are somewhat unclear — although they have the benefit that the top-ranked form will always be the form predicted, preventing discrepancies of the type we observe between speaker productions and ratings. Albright and Hayes (2003) evaluate their models by correlating their scores with both speaker rating data and production probabilities. Model scores are more often correlated to production probabilities (Hahn and Nakisa, 2000; Kirov and Cotterell, 2018), although whether this comparison is valid appears to depend on how one interprets the model scores, which is far from resolved in the case of ED models. Corkery et al. (2019) correlate individual model scores with speaker production probabilities, but note that this varies highly across models trained with separate random seeds. They observe as well that the model tends to place very high probability on the top-ranked form and relatively low scores on all other forms, leading to high instability in lower-ranked forms — individual models will often agree on the top-ranked form, and disagree in terms of the form in second place. In response to the difficulty of interpreting model scores, they instead aggregated production probabilities across models and compare their correlations with aggregated production probabilities of speakers, similar to the approach pursued in this chapter.

Figure 5.3 shows the distribution of predicted plural classes by rank for training,

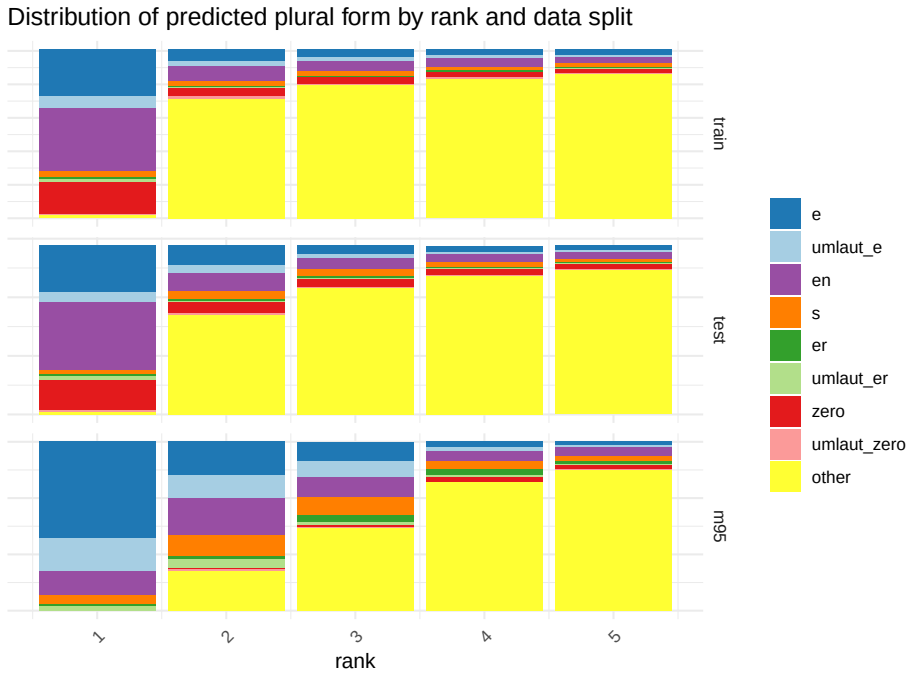

Figure 5.3: Distribution of predicted plural class by rank for train, test, and M95 (wug) data.

test, and wug data sets[3]. The disparity between the top-ranked outputs and lower-ranked forms is immediately obvious: while the top-ranked forms tend to correspond to an identifiable plural class, around 70% of second-ranked outputs are unattested 'other' forms, often involving some form of stem modification that does not typically figure into Germam plural formation (some examples from the training data: *Aachener* → *Aacherer, Flegel* → *Flegnle, Vorschuss* → *Vorscöüsse*). The German ED model also displays high concentration of probability mass on the top-ranked form, and relatively small differences in the scores at lower ranks. On average, when comparing within forms of the same rank, plausible outputs which conform to one of the established plural classes receive slightly higher scores than implausible 'other' outputs — but the differences are dwarfed by the gulf between the top-ranked form (average score -0.002) and the lower-ranked forms (average score -14, -15.2, -15.8, and -16.2 for ranks 2–5).

When considering the English past tense data, it is difficult to make broad claims about plausibility of lower-ranked model predictions — partly because the strong push toward the regular form /-d/ concentrates probability to a degree that might be desirable give its predominance, and partly because irregular forms are highly idiosyncratic; for

[3]Interestingly, the wug data appears to have a larger proportion of legitimate plural forms in its lower-ranked outputs. This could be partly due to the fact that the wug data comprised very short, monosyllabic words — observationally, longer words were more likely to yield distorted predictions, with erroneous modifications of the stem — but further investigation would be needed to confirm.

this reason, Corkery et al. (2019) calculate a recall metric based on whether the one or two irregular forms proposed by Albright and Hayes appear in the model's top 5 ranked predictions. In contrast, the German plural system offers a range of acceptable plural forms, and Figure 5.3 shows that the model's lower-ranked predictions often eschew plausible forms entirely in favor of spurious outputs. This observation further calls into question the suitability of ED model scores as approximations of speaker behavior: it is unclear that correlation between model scores and production probabilities is an informative metric if the scores barely distinguish between acceptable and wholly implausible outputs. The German speaker production data suggests that virtually all of the plural classes are at least somewhat acceptable for each word in the M95 wug stimuli; given this, it is also unclear whether the model's concentrated probability mass on the top-ranked form reflects an appropriate inductive bias for modeling cognition.

5.3 Discussion

Overall, the model's productions appear to be driven by similarity to items observed in the training set to a much greater extent than those of German speakers, an observation that is consistent with the criticisms leveled at connectionist modeling in the 90s, when this German plurals became a contentious topic within the broader cognitive debate. The powerful capacities of the encoder-decoder with attention seem able to approximate the overall performance of speakers on the M95 wug stimuli, as seen in the correlated production probabilities for /-s/ and /-e/. However, certain acute differences leave open the question of how and under which circumstances they may act as cognitive models; the failure to generalize /-en/ in particular appears to reflect a lack of appropriate learning mechanisms. It's possible that better task design (e.g. jointly learning mappings from the entire inflectional paradigm, which may reduce unattested errors such as first-letter stem alteration) and more information in the data (e.g. determiners, which would provide cues to headedness, gender, and other relevant factors) could lead to better inductive biases. At the very least, as a starting point, we can note that random initialization with the same hyperparameters does not at all approximate the diversity observed across human behavior — insofar as this is an important consideration in cognitive modeling, we may wish to consider how that range could be realized. In addition, the tendency for implausible outputs to dominate the model's second- and third-ranked predictions calls into question whether model scores can be considered suitable analogues for cognitive processes.

Chapter 6

Conclusions

This thesis has taken the claims of Marcus et al. (1995) as the starting point to pursue two research questions focused on the inflectional morphology of German number.

RQ1. Do German speakers treat /-s/ as the default plural for novel nouns?

The survey results presented in Chapter 3 show, to a first approximation, that all major plural classes in German have at least *someone* out there willing to extend them to novel nouns. /-(e)n/, /-e/, and /-s/, all of which have been proposed as productive plural suffixes in the literature (e.g. Indefrey, 1999; Zaretsky and Lange, 2016; Yang, 2016), are extended to the wug stimuli roughly in proportion to their type frequency in the corpus.[1] Even the plural suffixes /-er/ and the zero plural, which are widely considered nonproductive classes, together accounted for 15% of produced forms in the survey. The high degree of variability observed in individual speaker preferences seems relevant to future attempts at cognitive modeling for this inflectional system.

The other key outcome of the survey is that the evidence for /-s/ as default plural is decidedly weak. Speaker production of /-s/ for the M95 wug stimuli was around 5%, roughly proportional to the overall type frequency of the suffix, and in the rating data /-s/ ranked fifth out of the eight plural classes on average. These production probabilities for /-s/ are fairly low relative to other production wug tests (Köpcke, 1988; Zaretsky and Lange, 2016), and the ratings are far lower than those found by Marcus et al. (1995), raising the possibility that the suffix /-s/ is becoming less preferred over time. Beyond the overall distribution, a core empirical claim for /-s/ as default in the original M95 rating data was the preference for /-s/ in non-rhymes relative to rhymes; Clahsen (1999a) emphasizes that it was the only plural class that showed this pattern,

[1] Modulo the relative increase for /-e/ in production, likely influenced by the study design decision to assign the neuter article *das* for all stimuli.

confirming its singular status as regular under the dual-route model. Our survey results in production and rating both reproduce the /-s/ preference for non-rhymes (6% vs. 4% production probability), but find that /-(e)n/ is also preferred for non-rhymes relative to rhymes (35% vs. 25%), consistent with Zaretsky and Lange (2016). Although /-(e)n/ has been analyzed as a default plural for feminine nouns, the stimuli in our experiment were presented as neuter, so grammatical gender is unlikely to account for its prevalence. Overall, the evidence for /-(e)n/ as default plural is equal to, if not stronger than, the evidence for /-s/ default in our data — implying that a dual-route model which posits a single regular form cannot adequately account for these findings.

RQ2. Do modern neural networks learn the correct generalizations for the German plural system?

Kirov and Cotterell (2018) argue that modern neural networks have the capacity to inform cognitive modeling, as they learn the correct inductive biases to model speaker behavior for English past tense inflection. We use their proposed encoder-decoder architecture to evaluate whether this claim holds for the more complex system of German number inflection.

On a held-out test set of existing German nouns, the model achieves 87% accuracy overall, but performs around 50% on the /-s/ class as well as on the long tail of 'other', truly irregular plurals. Even though the minority-default analysis of /-s/ may not characterize speaker behavior on wug tests, it does appear to reflect the diverse range of phonological environments in which /-s/ occurs as a plural suffix — and this dispersion appears to pose a challenge for the model, as Marcus et al. (1995) anticipated. The model also shows variable ability to represent meaningful linguistic categories such as noun headedness. Once the plural class of a derivational head (e.g. the noun *Kind+/-er/* or the suffix *-ung+/-en/*) is learned, the plural class of all nouns sharing that head is fully predictable; however, it is unclear whether the model effectively learns and exploits these regularities.

On the M95 wug stimuli, the encoder-decoder predictions are broadly consistent with speaker behavior: both produce /-e/ most frequently, and aggregated production probabilities for individual items show a relatively high correlation around $\rho = .6$ between model and speaker. Model outputs also exhibit the non-rhyme preference for /-s/, which could indicate that the network has successfully learned the postulated inductive bias for phonologically atypical nouns. Nonetheless, certain other patterns call the model's performance into question. Most notably, it fails to generalize /-en/ to the extent speakers do — and its /-s/ generalization to non-rhymes may in fact be driven by

a serendipitous rhyming item in the training set: the noun *Generalstreik*, which takes /-s/ plural and rhymes with three of the twelve non-rhyme stimuli. Beyond pure performance, other aspects of the model's behavior differ from speakers in ways that are potentially relevant for cognitive research: individual items drive model predictions to a greater extent than speakers, variability across model simulations with different random seeds is much lower than variability across individual speakers, and the slew of implausible plural forms in the second and third rank of predicted outputs suggests that encoder-decoder scores do not necessarily reflect what speakers would consider acceptable.

It is possible that some of these differences reflect limitations of the current study. For example, grammatical gender interacts with plural class and provides a strong cue to headedness, but was absent from the training and test data; the model may have been capable of more accurate generalization with access to this information. However, the larger problem is one of definition: what are the desiderata for a model aiming to inform cognitive science? The debate around inflectional morphology since Rumelhart and McClelland (1986) has often turned on the definition of what constitutes a rule, a framing which has driven a great deal of scientific investigation but may also obscure productive modes of inquiry combining the two perspectives (Seidenberg and Plaut, 2014; Pater, 2019). Kirov and Cotterell usefully refocus the debate away from the rule-pattern opposition, and toward generalization behavior more broadly; nonetheless, the question of *how* models generalize drove much of the rule-pattern debate, and exactly those questions resurface in consideration of the German plural system examined in this study. Kirov and Cotterell point to the high performance of the encoder-decoder and its tendency to overregularize the /-d/ inflection as evidence that it has learned human-like inductive biases; Corkery et al. (2019) highlight the variable correlation with past tense wug tests across random seeds, and note the instability of the second-ranked outputs, as reasons to question the model's suitability. In this study, the variability of individual speaker preferences across plural classes suggests that instability may in fact be a *desirable* property of a cognitively plausible model — but the instability should presumably mirror human variation in meaningful ways, so the instability of the model's second-ranked implausible German plural forms likely does not tell us much about the plural formation processes of German speakers. Furthermore, the model's over-reliance on particular items and difficulty correctly generalizing /-s/ to the test set suggests that some of the criticisms Pinker and Prince (1988) originally leveled against connectionist modeling maintain their validity today.

High accuracy on morphological transduction of unseen test data is a promising start, but is it enough? If we wish to gain cognitive insights from modern neural networks, we need to specify *which* inductive biases are relevant to these lines of inquiry. Fortunately, the success of recent developments in deep learning allow us to develop powerful models with which evaluate these problems. In the case of the German plural system, these questions largely remain open.

6.1 Future work

There are many directions for future work on modeling German number inflection. A natural next step would simply extend the setup described in this paper to new datasets which might provide additional insight. Köpcke (1988) and Hahn and Nakisa (2000) both report results from other wug tests on German speakers, using more diverse stimuli. For example, Köpcke collected speaker productions for derived novel words with existing suffixes such as *-lein*; evaluating model predictions for these stimuli could provide more insight into the model's representation of headedness. Nakisa and Hahn (1996) also describe testing a neural network on an artificial language designed to represent the theoretical distribution of the default rule — such an approach could clarify the capacity of the model to represent some default rule independent of the particular case of German number. The model could also simply be given more training data: where UniMorph has around eleven thousand German nouns, a more recent dataset scraped from Wiktionary[2] contains eighty-three thousand.

Another obvious next step would be the addition of grammatical gender, an important cue to plural class, in training and test data. The larger Wiktionary dataset also contains grammatical gender information. Preliminary experiments on this data suggest that the model learns a similar mixture of desirable and undesirable inductive biases: it is highly responsive to grammatical gender, much like German speakers, but it also learns to overgeneralize /-e/ on nonfeminine nouns to an even greater extent. In addition to grammatical gender, the larger dataset contains forms inflected for number across different cases. Jointly learning an entire inflectional paradigm, with multiple inflected forms for each word, may also support better model generalization, as found by Kirov and Cotterell (2018) in their second experiment.

Further experiments might investigate alternative neural architectures beyond encoder-decoder. Hahn and Baroni (2019) find some evidence that character-level language

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

models learn to generalize German number inflection; controlled wug testing of these models could help ascertain what sort of inductive biases are learned under their relatively unsupervised training process. Another line of inquiry might flip the problem on its head and use neural models to identify minimal pairs of stimuli which lie on some decision boundary — e.g. a nonce word with a high score for one plural class, but which flips to some other plural class when one phoneme is changed — and evaluate the degree to which speaker preferences in wug tests reflect these predicted differences.

Bibliography

Albright, A. and Hayes, B. (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition*, 90(2):119–161.

Baayen, R. H., Piepenbrock, R., and van H, R. (1993). The CELEX lexical data base on CD-ROM.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv: 1409.0473.

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.

Berko, J. (1958). The Child's Learning of English Morphology. *WORD*, 14(2-3):150–177.

Bittner, D. (2000). Gender classification and the inflectional system of German nouns. *Trends in linguistics studies and monographs*, 124:1–24.

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.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, 10(5):425–455.

Clahsen, H. (1999a). The dual nature of the language faculty. *Behavioral and Brain Sciences*, 22(6):1046–1055.

Clahsen, H. (1999b). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences*, 22(6):991–1013.

Clahsen, H., Rothweiler, M., Woest, A., and Marcus, G. F. (1992). Regular and irregular inflection in the acquisition of German noun plurals. *Cognition*, 45(3):225–255.

Clark, A. (2002). Memory-Based Learning of Morphology with Stochastic Transducers. pages 513–520.

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*, pages 3868–3877, Florence, Italy. Association for Computational Linguistics.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., McCarthy, A. D., Kann, K., Mielke, S., Nicolai, G., Silfverberg, M., Yarowsky, D., Eisner, J., and Hulden, M. (2018). The CoNLL-SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. page 27.

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*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Dressler, W. U. (1999). Why collapse morphological concepts? *Behavioral and Brain Sciences*, 22(6):1021–1021.

Dreyer, M. and Eisner, J. (2011). Discovering Morphological Paradigms from Plain Text Using a Dirichlet Process Mixture Model. pages 616–627.

Durrett, G. and DeNero, J. (2013). Supervised Learning of Complete Morphological Paradigms. pages 1185–1195.

Elgersma, D. and Houseman, P. (1999). Optimality Theory and Natural Morphology: An Analysis of German Plural Formation*. *Folia Linguistica*, 33(3-4).

Elsen, H. (2002). The acquisition of German plurals. In *Morphology 2000: selected papers from the 9th Morphology Meeting, Vienna, 24-28 February 2000*, number v. 218 in Amsterdam studies in the theory and history of linguistic science, pages 117–127. J. Benjamins, Amsterdam ; Philadelphia.

Gandhi, K. and Lake, B. M. (2019). Mutual exclusivity as a challenge for neural networks. *arXiv:1906.10197 [cs]*. arXiv: 1906.10197.

Hahn, M. and Baroni, M. (2019). Tabula nearly rasa: Probing the Linguistic Knowledge of Character-Level Neural Language Models Trained on Unsegmented Text. *arXiv:1906.07285 [cs]*. arXiv: 1906.07285.

Hahn, U. and Nakisa, R. C. (2000). German Inflection: Single Route or Dual Route? *Cognitive Psychology*, 41(4):313–360.

Halle, M. (1973). Prolegomena to a Theory of Word Formation. *Linguistic Inquiry*, 4(1):3–16.

Haspelmath, M. and Sims, A. D. (2010). *Understanding morphology.* Routledge, London.

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)*, volume 7, pages 136–153. ERIC.

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)*, pages 555–560, Berlin, Germany. Association for Computational Linguistics.

Kawahara, S. (2016). Psycholinguistic Methodology in Phonological Research. Technical report, Oxford University Press. type: dataset.

Kirov, C. and Cotterell, R. (2018). Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker and Prince (1988) and the Past Tense Debate. *arXiv:1807.04783 [cs]*. arXiv: 1807.04783.

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)*, pages 3121–3126, Portorož, Slovenia. European Language Resources Association (ELRA).

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., and Rush, A. M. (2018). OpenNMT: Neural Machine Translation Toolkit. *arXiv:1805.11462 [cs]*. arXiv: 1805.11462.

Köpcke, K.-M. (1988). Schemas in German plural formation. *Lingua*, 74(4):303–335.

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.

Laaha, S. (2011). Sonority, gender and the impact of suffix predictability on the acquisition of German noun plurals. *Language, Interaction and Acquisition*, 2(1):82–100.

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.

Markman, E. M. and Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20(2):121–157.

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.

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*.

Pinker, S. (1999). *Words and rules: The ingredients of language*. Basic Books, New York.

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.

Pinker, S. and Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, 6(11):456–463.

Plunkett, K. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition*, 48(1):21–69.

Prasada, S. and Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, 8(1):1–56.

Pulvermüller, F. (1999). Please mind the brain, and brain the mind! *Behavioral and Brain Sciences*, 22(6):1035–1036.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rumelhart, D. E. and McClelland, J. (1986). On Learning the Past Tenses of English Verbs. page 56.

Seidenberg, M. S. and McClelland, J. L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. *Psychological Review*, 96(4):523–568.

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.

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.

Stemberger, J. P. (1999). Frequency determines defaults in German: Default perfect-t versus irregular plural-s. *Behavioral and Brain Sciences*, 22(6):1040–1041.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, pages 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)*, pages 674–680, Beijing, China. Association for Computational Linguistics.

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.

Wunderlich, D. (1999). German noun plural reconsidered. *Behavioral and Brain Sciences*, 22(6):1044–1045.

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 153 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, pages 153–178. University of Bamberg Press, Bamberg.

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.

Appendix A

Experimental Materials

A.1 Stimuli

Table A.1 provides the complete list of nouns used in the experiment.

Rhymes	Non-rhymes
Bral	Bnaupf
Kach	Bneik
Klot	Bnöhk
Mur	Fnahf
Nuhl	Fneik
Pind	Fnöhk
Pisch	Plaupf
Pund	Pleik
Raun	Pläk
Spand	Pnähf
Spert	Pröng
Vag	Snauk

Table A.1: Experimental stimuli (Marcus et al., 1995).

A.2 Presentation

All stimuli were presented in their singular form as neuter gender, i.e. they were preceded by the article *Das*.

Figure A.1: Example presentation of two items during the production task

Participants first completed a production task in which they were prompted to produce a plural form for all stimuli by typing into the box below the plural prompt *Die*. Figure A.1 gives an example of the presentation of two items during the production task.

Participants then completed a rating task in which they were prompted to rate the acceptability of plural forms for all stimuli. Figure A.2 gives an example of the presentation of an item during the rating task.

Note that this item includes an 'attention check', as one of the forms to be rated simply indicates the correct option to select (in this case, 'Sehr gut'). Participants who did not answer these attention checks correctly were excluded from analysis.

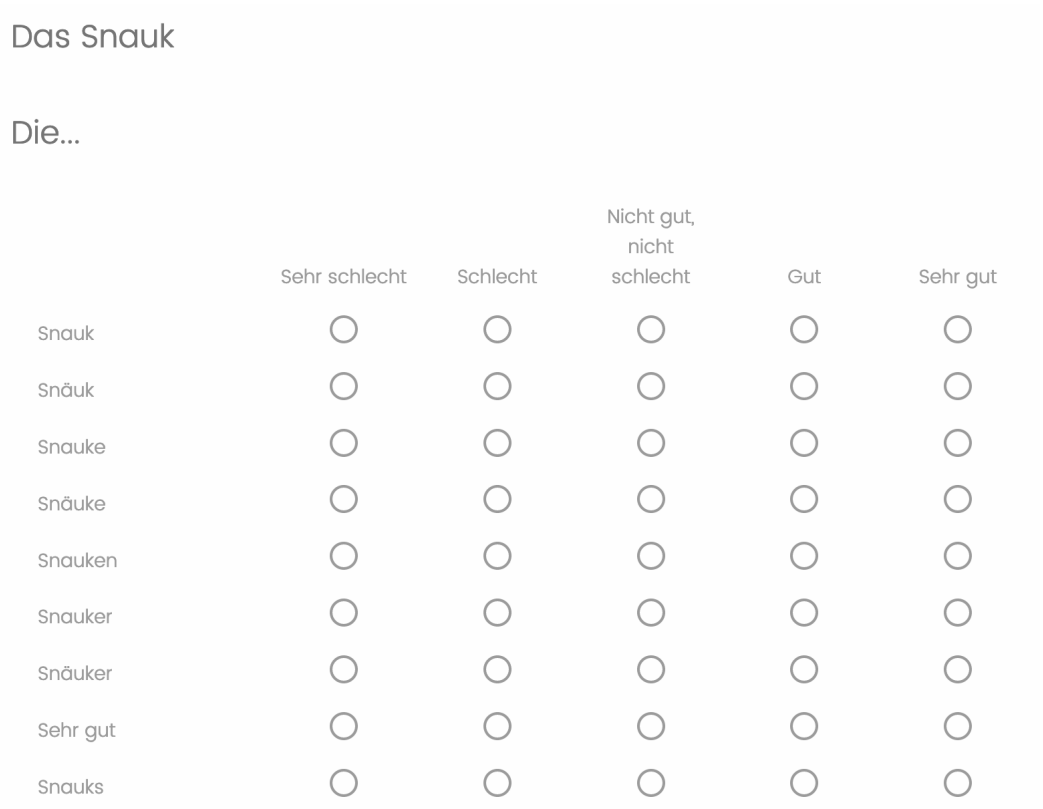

Figure A.2: Example presentation of an item during the rating task