Simulating the influence of semantics on German noun inflection

Maria Heitmeier (B166245)

Master of Science
Cognitive Science
School of Informatics
University of Edinburgh
2020

Abstract

Inflection describes the process of altering words to express functions such as tense, grammatical gender or number (Crystal, 2009). Recently, McCurdy (2019) used an architecture from Natural Language Processing, the Encoder-Decoder model (Sutskever et al., 2014), as a cognitive model of German pluralisation and compared its predictions to those of German speakers on a set of made-up wug words. While the productions of models and speakers were broadly correlated, the model also showed some clear failure modes. Crucially, these models rely only on the form ("phonology") of the words to predict new forms, but research suggests that the meaning of words ("semantics") might be predictive for inflection as well (Ramscar, 2002; Baayen and del Prado Martín, 2005; Williams et al., 2020). We therefore hypothesised that a) speakers' plural productions are influenced by semantics and b) that ED models' predictions would be more speaker-like with the inclusion of semantic information. We conducted an experiment with German speakers, asking them to produce and rate various plural forms of made-up German words in a semantic context. The effects of semantics remained somewhat unclear, as they only partly matched the expected effects in corpus data. The influence of semantics on the models differed from that on speaker productions, suggesting that the ED models make use of the semantic information in a different way than speakers. Nevertheless, the models showed a slightly higher correlation to the speaker data than previous work and than models without semantic information. Possibly, semantic information leads to more diversity in the model predictions, thus improving its fit to the speaker data, whose variability is a key characteristic (also McCurdy (2019); Zaretsky et al. (2013)). Therefore, semantics were not found to have a clear influence, and our results imply that what drives inflection in German is still largely unclear.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Maria Heitmeier (B166245)*)

Acknowledgements

I would like to thank my supervisor, Stella Frank, for her advice and support throughout the project. I am grateful to the Wrap Gang for an awesome winter and what would have been an even better summer, my family for their company during lockdown, and Marius Hobbhahn, for his continued support.

Table of Contents

Chapter 1

Introduction

Many languages mark their words for certain functions, such as tense, grammatical gender or number, a process known as inflection (Crystal, 2009), creating a complex system of different word forms. Humans encounter new words during their entire lifetime (Blevins et al., 2017) and they are able to use them in a grammatically correct manner, generalising to new inflectional forms without effort. This is generally known as the *paradigm-cell filling problem* (Ackerman et al., 2009). In the late 1980s a debate was sparked focusing mostly on the mechanisms behind the English past tense, known as the "past tense debate" (e.g. Pinker and Ullman, 2002; Seidenberg and Plaut, 2014). The main point of dispute was whether the English past tense is produced using two mechanisms, a rule-like one for regular and retrieval from memory for irregular verbs (Pinker and Prince, 1988), or with a single mechanism, usually a connectionist network as first proposed by Rumelhart and McClelland (1986).

Since then, the debate has moved somewhat away from a simple distinction between single- and dual-route models. For example, Albright and Hayes (2003) suggested a stochastic rule-based model, and subsequently the debate started to shift slightly to comparing analogical or connectionist to rule-based models (Albright and Hayes, 2003; Kirov and Cotterell, 2018; Corkery et al., 2019). Additionally, increasingly more morphologically complex languages such as German, Maltese or Irish have been explored (McCurdy, 2019; Malouf, 2017; Cardillo et al., 2018), heightening the interest in what really drives inflection, instead of focusing on only two models.

In recent years, Natural Language Processing (NLP) has developed a multitude of successful Neural Network architectures - the next generation of connectionist models. Kirov and Cotterell (2018) thus suggested to use such a modern architecture, the Encoder-Decoder (ED) model, as a cognitive model of inflection. ED architectures have been used successfully in e.g. Machine Translation (e.g. Sutskever et al., 2014) or morphology modelling (e.g. Cotterell et al., 2016). Kirov and Cotterell (2018) reported success in modelling the English past tense. They also tested their model on so-called wug words - invented words designed to explore how humans generalise to new word forms (Berko, 1958) - and reported a reasonably high correlation between their model predictions and speaker productions. However, Corkery et al. (2019) found that when running multiple random simulations, the correlation between model and speaker data was highly variable. Subsequently, McCurdy (2019); McCurdy et al. (2020) used the architecture to model the German plural system, which - contrary to the simple English system - has three grammatical genders and eight plural classes with none being prevalent overall. This model, too, was only broadly correlated to the productions they collected from German speakers. They also found their ED models to show considerably less variability than German speakers.

These models have in common that they feed only the phonology/orthography of the singular forms as input to the model. However, research suggests that the meaning of a word (its semantics) can be informative of its inflection class as well. For example, Ramscar (2002) found for English that participants could be semantically primed to use regular or irregular past tense forms. Baayen and del Prado Martín (2005) showed that distributional semantics vary systematically between regular and irregular verbs in various Germanic languages. And finally, Williams et al. (2020) found that semantics are predictive for German inflectional classes.

We therefore developed two hypotheses:

(1) German speakers are influenced in their selection of plural classes by the semantics of words.

(2) Feeding semantic information to an Encoder-Decoder model of German plural inflection improves the fit between model predictions and productions of speakers.

In order to test (1) we conducted a wug word experiment, making use of an observation by Gaeta (2008): German masculine nouns which describe a person are more likely to take an *-n* plural than ones describing an object. We thus expected speakers to show a higher probability of using *-n* plurals if the wug word they were presented with described a person than when it described an object. We found that German speakers do not show this behaviour: they were only slightly more likely to use *-n* plurals in the *person* condition, and the difference was not statistically significant. On the other hand, they unexpectedly used fewer *-e* plurals in the *person* condition, a tendency which could also be found in corpus data. Taken together, this suggests that we cannot completely reject the hypothesis that German pluralisation is influenced by semantics - but that this influence might be part of a more complex system which has not been (fully) captured in the present study.

Because our hypothesis was only applicable to a specific set of words, we created a new set of wug words. This led to unexpected findings: the German speakers in our experiment did not use *-s* words with more unfamiliar words, contrary to results of Marcus et al. (1995); McCurdy (2019). This implies that findings on a single set of wug words are not necessarily generalisable to other words.

In order to test (2), we explored various techniques of integrating semantics in the model. The most successful method was to include one of six semantic categories as semantic information. Such models did not reach considerably higher accuracy on a held-out test set (89.4%) than reference models trained without semantic information (89.2%). The effect of semantic category on the use of plural classes pointed into different directions than in the speaker data. The correlation between model and speaker data was slightly higher than in previous work and than compared to the reference models - but we were unable to conclusively link this to the effect of semantics. Our results suggest that the inclusion of further cues to plural class might improve the variability both within and across words and models, a key characteristic of German speakers' plural productions of wug words. Further research is necessary here, however, since a) the training regime was slightly changed compared to previous work and b) the number of reference models should be increased.

The remainder of this thesis is structured as follows: Chapter 2 gives a closer overview over the German plural system, wug studies, attempts at modelling the German plural system computationally and evidence for the influence of semantics on inflection. Chapter 3 reports results of an experiment conducted with German speakers designed to test hypothesis (1). Chapter 4 describes our attempts at including semantic information into the the model, results on held-out test data and an initial overview of its predictions on wug data, thus exploring hypothesis (2). In Chapter 5 we analyse and discuss the results from both speaker and model experiments to draw conclusions on hypothesis (2): does the model provided with semantic information show a better fit to speaker data? Finally, Chapter 6 summarises our conclusions and provides a brief overview over limitations and potential future work.

Chapter 2

Background

In this section we will give a short overview over the German plural system, previous results from wug word experiments and attempts to model their results. Finally, we will argue why we propose including semantics into McCurdy (2019)'s model architecture in order to improve the model's fit to data collected by German speakers.

2.1 The case of the German plural system

2.1.1 German plural system

The German plural system is generally divided into 8 different classes (Köpcke (1988); Schulz and Griesbach (1981); though e.g. Hahn and Nakisa (2000) count as many as 60). They either take *-n, -e, -s* or *-er* as suffix, or do not change their ending (henceforth -∅). Additionally, words in the *-e, -er* and -∅ classes can turn their root vowel into an umlaut (ä, ö, ü), henceforth *uml.* Apart from some subregularities there are no clear rules governing the relation between a word's singular form and its plural class (e.g. Köpcke, 1988). None of the classes is particularly prevalent. Overall, *-n* is the most frequent, followed by *(uml)-e* and *(uml)-∅*[1]. Moreover, a correlation between grammatical gender and plural class can be observed. The feminine class has a clear preference for *-n* (97%), while the distribution is more mixed for masculine and neuter nouns (see Table 2.1).

Because the German plural system is complex compared to the English past tense,

[1] Not all corpora and methods agree (Zaretsky and Lange, 2015). While the ranking reported by Gaeta (2008) and by Sonnenstuhl and Huth (2002) (estimated from CELEX, Baayen et al., 1995) agree, they differ in their estimated token frequency. For example, Gaeta (2008) reports a token frequency for the *(uml)-∅* plural of 9.5%, while Sonnenstuhl and Huth (2002) note 21%.

Gender	Ending	Example	Type / Token	Total
F	*-en*	Frau (woman), pl. Frauen	97.2% / 97.0%	53.5% / 50.0%
	uml-e	Hand (hand), pl. Hände	2.4% / 2.6%	
	uml-∅	Mutter (mother), pl. Mütter	0.4%/0.4%	
M/N	*(uml)-e*	Tag (day), pl. Tage	58.2% / 63.3%	38.8% / 47.2%
	(uml)-∅	Muster (pattern), pl. Muster	33.8.% / 19.7%	
	uml-er	Mann (man), pl. Männer	5.9% / 6.6%	
	-en	Auge (eye), pl. Augen	2.1% / 10.4%	
animate M	*-en*	Mensch (human), pl. Menschen		3.7% / 1.6%
-	*-s*	Opa (granddad), pl. Opas		2.6% / 0.9%

Table 2.1: Frequency of plural endings of German nouns in different genders. Table adapted from Gaeta (2008, p. 79) as in Heitmeier (2020)

it has sparked some debate about which model might best account for it (e.g. Köpcke, 1988; Marcus et al., 1995; Hahn and Nakisa, 2000; McCurdy, 2019). In particular, while the English past tense has an obvious default/regular rule (*-ed*), the German plural system does not. Marcus et al. (1995) proposed *-s* as a minority default rule, which has been contested by others (e.g. Hahn and Nakisa, 2000). Both sides have regularly claimed results from so-called Wug studies as supporting their arguments.

2.1.2 Wug studies

Wug studies are a popular method to investigate word inflection. First introduced by Berko (1958) they investigate how inflection is generalised to new, unknown words. Participants are presented made-up words (as for example 'wug') and are asked to inflect them. If speakers use generalisation rules these should be visible in the inflection pattern. One of the first wug studies with German adult speakers was conducted by Köpcke (1988). They found that in general the *-n* and *-s* classes were overgeneralised in words where the gender and ending did not dictate a single class (see Table 2.1), as compared to their general corpus distributions. Marcus et al. (1995) subsequently hypothesised that the *-s* class is used as the default by German speakers. They conducted an experiment where 24 wug words were presented in three different contexts: as names, as borrowings from other languages and as 'root' word, essentially using it like an existing German noun. Marcus et al. (1995) expected that the less a wug word is used like a normal German noun, the more likely it is to take its default plural. Additionally, half of the words had one or more rhymes in German, thus making pluralisation by analogy to other German nouns possible for this subset. They assumed that participants would be more likely to use *-s* plurals with the non-rhyme words, their lower familiarity triggering a default response. Participants were asked to rate all possible plural forms. Marcus et al. (1995) found that for words without a rhyme participants preferred *-s* plurals compared to non-rhyme words, thus supporting the claim that *-s* is used as a default by German speakers. However, while their hypothesis held overall, an effect of preferring *-s* plurals over others was only visible in the 'name' condition. This suggests an effect of context in how German speakers inflect nouns - possibly an indication of an influence of semantics (see below) (see also Hahn and Nakisa (2000) for a closer exploration of German plural production in names).

Marcus et al. (1995)'s results were contradicted by Zaretsky and Lange (2015) who in a production task with the same target words (but without the contexts) and 585 participants found that the *-e* class dominated for both rhyme and non-rhyme words. Finally, McCurdy (2019) ran an experiment with two tasks: first freely producing plural forms, then rating all possible forms. They found that the rating data differs widely from the production data. While speakers preferred the *-e* (38%) class in production, followed by *-n* (30%) and *-s* (5%), in the rating task they preferred the *-n* class (M 3.8) over *-e* (3.5) and *-s* (2.5). Moreover they reported a negative effect of rhyme on the use of the *-s* class. Variability across both participants and items was very high. Given the results from all previous studies, what exactly governs the use of plural classes in German nouns is still largely unexplained.

2.1.3 Computational Models

Cognitive models allow a researcher to explain observed data and make clear predictions for future observations (Lewandowsky and Farrell, 2010, ch. 1). Various computational models such as nearest neighbour or Generalised Context Model (Nosofsky, 1986) have been explored to model inflection in the past (Hahn and Nakisa, 2000). Most prominently, Rumelhart and McClelland (1986) proposed a simple fully connected feed-forward linear model (Kirov and Cotterell, 2018). However, it had distinct failure modes as observed by Pinker and Prince (1988), not the least among them that it performs poorly on held-out test data. A wealth of both neural and non-neural models was subsequently proposed for English (e.g. Plunkett and Juola, 1999; Albright and Hayes, 2003) as well as German (e.g. Nakisa and Hahn, 1996; Westermann and Miikkulainen, 1994), but their success was mixed (Kirov and Cotterell, 2018). Kirov and Cotterell (2018) suggested that the original architecture by Rumelhart and McClelland (1986) was simply not powerful enough to model English past tense inflection and subsequently used an Encoder-Decoder (ED) Neural Network, a state-of-the-art deep learning architecture used for tasks such as machine translation (Sutskever et al., 2014), and recently also for morphological paradigm completion (Cotterell et al., 2016). They reported not only high accuracy on unseen data (94.5%), but also found that the error pattern of the model was in general consistent with human speech errors. Furthermore, they tested their model on the 74 wug words by Albright and Hayes (2003). They found that the responses of their model correlated with the responses collected by Albright and Hayes (2003). However, Corkery et al. (2019) tried to replicate their results and found that the predictions and thus the correlation between model and speaker data vary widely depending on the initialisation of the model. Even when aggregating the data across multiple random simulations, correlation was low.

Consequently, McCurdy (2019) and McCurdy et al. (2020) trained an ED model on German plurals and compared its performance to German speaker data. Since Marcus et al. (1995) had claimed that German has a minority default plural class (*-s*) it was expected that the ED model might fail to generalise to a minority in the data. While they did find that the ED model did only use the *-s* plural in a minority of cases, their speaker data showed that German speakers do the same (see above). Still, the model data did not show a good fit to the human speaker data. It a) failed to use the *-n* class as much as speakers do and b) variability across models was much lower than across speakers. They concluded that ED models are perhaps simply not well suited as cognitive models of German noun inflection.

2.2 The influence of semantics

Classical models of inflection do not take into account semantics. For example, Pinker and Prince (1988, p. 113, sic) claim that “the specific semantic distinctions between, say, *ring* and *wring*, are hardly the basis for any real generalization”. Analogously, while the words “*hit, slap* and *strike*” are very similar in meaning, they each form their past tense differently (Seidenberg and Plaut, 2014, p. 1203). However, research suggests that there might be an influence of semantics after all. Ramscar (2002) showed their participants wug words that were introduced within a text either in the place of a regular or an irregular verb (without ever specifically mentioning the verb), and found that depending on the surrounding text and thus the regularity of the induced verb, participants tended to form the past tense of the wug words regularly or irregularly. Baayen and del Prado Martín (2005) reported a difference between the semantic clustering of regular and irregular verbs in multiple Germanic languages. Williams et al. (2020) used word embeddings to predict a word's plural class in German and Czech (controlling for grammatical gender) and found that the interaction of meaning and form consistently predicted plural class better than form alone. As regards wug words, Cassani et al. (2019) generated semantic embeddings from the phonology of wug words (compiled to either resemble verbs or nouns) by using a computational model which directly maps phonology to semantics. They found that the semantic embeddings were able to not only predict the word category, but also children's responses to an entity/action discrimination task. And finally, research by Chuang et al. (2020) suggests that speakers' lexical decision times on wug words can be predicted by their semantic embeddings, generated using the same method as Cassani et al. (2019). These results question whether wug words are really devoid of meaning as is generally assumed in wug studies (Chuang et al., 2020), and indicate that wug word inflection cannot accurately be modelled without the inclusion of semantics.

Neural Networks can accommodate semantics fairly easily (e.g. Hoeffner, 1992; Joanisse and Seidenberg, 1999) by adding it to the input which the models try to predict (plural) forms from. In order to pin-down the hypothesised effect of semantics, we make use of an easily observable example reported by Gaeta (2008): they observed that German masculine multiple syllable nouns describing a person (e.g. *der Ma-trose (sailor)*) are more likely to belong to the so-called weak inflection class than single syllable nouns describing something inanimate (e.g. *der Stein (stone)*). The weak inflection class is characterised by taking an *-n* plural and an *-n* singular genitive form. We assumed that this observation would also hold true for the entire *-n* plural class (and we confirmed this for corpus data, see Section 3.1). Within this study we will use Gaeta (2008)'s observation as a hypothesis to observe whether semantics influence how words are inflected. If semantics have an influence, human participants should be more likely to use an *-n* plural for nouns presented as describing a person than for nouns describing an object. If ED models serve as an accurate description of German speakers' behaviour in wug tasks, they should show an equal effect of semantics and the overall correlation between speaker and model data should increase.

Chapter 3

Wug study with German speakers

We aimed to test hypothesis (1), that semantics influences the plural productions of German speakers, by carrying out a wug study. In order to evaluate the results we made use of an observation by Gaeta (2008): a German noun is more likely to be assigned the *-n* plural class if it has multiple syllables and, crucially, describes a person, as opposed to an object. In the following, we first describe the steps taken to verify this observation. After a description of the materials and procedure of the experiment, we report and discuss its results.

3.1 Verifying Gaeta (2008)'s observation on corpus data

We used all nouns in a German corpus[1] scraped from Wiktionary[2], which could be translated to English and had an entry in WordNet (Fellbaum, 1998). From WordNet we extracted whether the noun described a person or one of five other semantic categories ('Semantic dataset', closer description of the process in Section 4.2). All nouns with masculine gender were used to run a logistic regression with presence or absence of the *-n* plural class as dependent variable and whether the word describes a person or not as predictor. Whether a word describes a person is highly predictive for the use of the *-n* plural ($p < 0.001$, full model in Appendix C). Gaeta (2008)'s statement is therefore true at least for corpus data. We did not investigate the additional claim that the *-n* plural is even more likely if the word has multiple syllables, because this hypothesis is nested in the first hypothesis which clearly holds.

[1] `https://github.com/gambolputty/german_nouns`

[2] `https://de.wiktionary.org/wiki/Wiktionary:Hauptseite`

3.2 Methods

3.2.1 Participants

We recruited 199 adult German native speakers on Prolific[3]. The answers of two participants had to be excluded because they failed an attention check (see Section 3.2.3). The remaining participants had a mean age of 30.61 (SD 10.23). 88.9% of participants received their primary education in Germany, 7.1% in Austria, 3.54% in Switzerland. Regional dialects vary for example in the gender they assign to some nouns, which might have an effect on their judgements (e.g. *butter* in Standard High German *die Butter*, in Bavarian *der Butter*). 51% of participants held a University degree. Participants having received higher education might have encountered a higher number of words previously unknown to them and might therefore respond differently to their use. Participants were compensated with a minimum of 5£ per hour.

3.2.2 Materials

Since Gaeta (2008) originally applied their observation to multiple-syllable words and because the wug words created by Marcus et al. (1995) have been shown to be problematic (one of the alleged non-rhyme words actually has a rhyme (McCurdy, 2019), many of the words have the same ending (e.g. *Bneik, Fneik, Pleik*) or show unusual orthography (Zaretsky and Lange, 2015)), we created our own set of wug words. They were compiled by making use of *Wuggy* (Keuleers and Brysbaert, 2010), a pseudoword generator. They were then manipulated so that endings either had no rhymes in a corpus from German Wiktionary[4] but were phonologically plausible (25% of items, e.g. *Jakaselb*, *Rerofept*) or had rhymes which occurred with different plural endings in German (75%, e.g. *Katulee (der See/die Seen, der Tee/die Tees)* or *Filast (der Mast/die Masten, der Gast/die Gäste)*), to make sure that no single ending was especially salient. Note that since rhyme was not the focus of this investigation, it was not balanced across the other factors. Words were additionally manipulated to ensure that the words were reasonably diverse, i.e. have a range of different starting letters. Of the overall 24 presented wug words, half have two, the other half three syllables. All words can be found in Appendix A.

For the rating task we generated up to eight possible plural forms per wug word[5].

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

[4] `https://github.com/karoly-varasdi/de-wiktionary-parser`

[5] Due to experimenter error, three of four plural forms of the word *Walgimirz* included a typo, with *l*

In words with multiple vowels we opted to form the umlaut of the last vowel in the word in order to reduce the number of possible plural forms (and thus not exhaust participants with up to 14 different plurals).[6] If the last vowel did not have an umlaut, we did not provide any umlaut forms.

Images were taken from ImageNet (Deng et al., 2009) and the ESP game data set (Von Ahn and Dabbish, 2005). We selected images depicting especially peculiar persons or objects (each 50%), to make introducing a new category name plausible and to reasonably assume participants would not know a category name for the depicted person/object. We aimed images to show persons who are adult male European/Caucasian and do not showcase a particular culture (to not confound peculiarity with culture). For images of objects we made sure that they are reasonably diverse, i.e. made from different materials (e.g. wood, metal, glass) and had different purposes (e.g. tool, musical instrument, vessel). For all images we ensured that a single person/object is depicted, if necessary by cropping the image.

The context sentences describe the shown image and include the target word three times. Since the plural of German (family) names is often different to that of regular nouns (cf. *Familie Bauer/die Bauers* vs. *der Bauer/die Bauern*, Marcus et al. (1995); Hahn and Nakisa (2000)), the target word is introduced as a term for a person/object **category** as opposed to a name for a **single** person/object, by using German pronouns introducing categories (e.g. *jeder (each)*). This also enabled us to introduce the word as grammatically masculine. These pronouns were also presented with every plural form in the rating task to make the plural more salient. The last sentence was designed to elicit the plural form of the target word, with the place of the word marked with a blank. The description sentences' nouns belong to a range of different plural classes,

and *g* being swapped (i.e. *Waglimirz*), and one of *Zackabat*'s plural forms was presented as *Zackbäter*, omitting a letter. While we were reasonably sure that participants generally did not notice, we nevertheless ran the analysis regarding the rating data both with and without these items.

[6] This decision is debatable. The root vowel is a lexical feature of a German word, and thus cannot be determined in wug words. Additionally, even though our goal was to use monomorphemic words, not all plural classes occur with monomorphemic multi-syllable words. For example, the *uml-e* class mostly occurs with compositional multi-syllable words (e.g. *Anlass/Anlässe*) (Schulz and Griesbach, 1981)). Here, the root vowel is the vowel of the last syllable/part of the word, but this is not the case in some monomorpehmic two-syllable words (e.g. *Apfel/Äpfel*). However, there seem to be hardly any monomorphemic three-syllable words with an umlaut at all (the number of monomorphemic three-syllable words is very low in German in general). We therefore assumed that *uml* in the initial syllables in especially three-syllable words would sound strange to Germans, and used the final vowel as root vowel, risking participants interpreting them as non-monomorphemic (given the low number of three-syllable monomorphemic words in German this was likely anyway). Our study results confirmed this intuition: Participants generally preferred to change the last vowel in a word (83% of all umlaut changes), in only 14% did they change the first vowel.

in order to avoid priming effects[7]. For illustration, the following example for sentences describing an object is translated to English (**wug** was randomly replaced with one of the wug words):

> A **wug** is standing in the dark. When it is completely dark, this **wug** is able to glow. This **wug** can be combined, further ___ can be stuck on top.

Analogously, a sentence describing a person:

> A **wug** is standing at the platform. Each **wug** prefers to wear a hat made from flowers on sundays. This **wug** and other ___ are currently on their way to visit friends.

All sentences as well as an example of how sentences and images were presented can be found in Appendices A and B, the Project Materials include all images used.

For practice trials we used three two- and three three-syllable words describing objects or animals from six different plural classes (*∅, -e, -en, uml, uml-e, -s*). Two classes are missing, because German has virtually no multiple-syllable, non-compositional nouns with an *(uml)-er* ending. Matching images were taken from ImageNet. Sentences were selected as described above, containing valid information about the selected words.

The experiment was implemented in Qualtrics[8]. Ethics approval was granted by the University of Edinburgh School of Informatics Ethics Committee (RT 2019/38242).

3.2.3 Procedure

Previous work has shown that participants tend to behave differently in production and rating tasks (McCurdy, 2019). We therefore opted to follow McCurdy (2019) and include both tasks in our study. The experiment had three parts: part I was designed to familiarise participants with the task. They were asked to produce the plurals of the six practice words with accompanying description sentences as described above. Part II was again a production task with all 24 wug words. Participants were randomly assigned one of two conditions. In condition A, half of the words were presented as a person category, the other half as an object category. In condition B, groups were swapped, to ensure that all words were presented with persons half of the time, and with objects the other half of the time. Each half included six two- and six three-syllable words. The order of presentation as well as the mapping between word and

[7] We verified this by checking for priming effects in the final results, none were found (Section 3.3).

[8] `https://www.qualtrics.com`

sentence/image was randomised. In part III, participants were asked to rate all plural forms (ranging from 4 to 8 possible forms) of 8 out of the 24 wug words from part II on a 5-point scale from ’Very bad’ to ’Very good’ (see Appendix B). The words were chosen randomly, making sure that they include four person and four object images/sentences, each again with half being two-syllable, the other half three-syllable words. While the order of wug words was randomised, the order of possible plural forms was constant (as in McCurdy (2019)). One of the trials included an attention check where participants were asked to select a given rating (see Appendix B).

After finishing all three parts, participants were asked for their age, whether they held a University degree and which country and state they had received their primary education in.

3.3 Results

3.3.1 Production task

Overall. The distribution of plural classes in the production task is consistent with previous work. Participants overall favoured the *-e* ending (43%), followed by *-n* (23%) and *-s* (10%). This is broadly consistent with McCurdy (2019) (38%, 30%, 5%) and almost mirrors the data by Zaretsky and Lange (2015) (46%, 24%, 12%) (the distribution of all classes is depicted in Figure 5.1).

The distribution of classes does not agree with corpus data. It differs from the overall distribution of plural classes in the Wiktionary corpus of German nouns (see Section 4.2, ‘semantic dataset’) ($\chi^2 = 7616, df = 81, p = 0.0$). This still holds when only using masculine person and object words from the corpus data ($\chi^2 = 7935, df = 64, p = 0.0$). When splitting the data into two semantic categories, thus each plural class appearing twice, there is still a clear difference ($\chi^2 = 18708, df = 289, p = 0.0$). The main difference is that participants in the experiment use the -∅ plural much less frequently (10%) than it is used in the part of the corpus with masculine grammatical gender (34%). Instead, they use *-e* (43%) and *-n* (23%) plurals more frequently (corpus 23%, 17%). The use of umlauts is also consistently higher in the corpus (17%) than in the production data (2%). Clearly, plural productions in German cannot easily be explained by distributions in corpus data.

Statistical model. In order to assess the influence of demographic factors and independent variables we used a Generalised Linear Mixed Effects Model (GLMM) (fol-

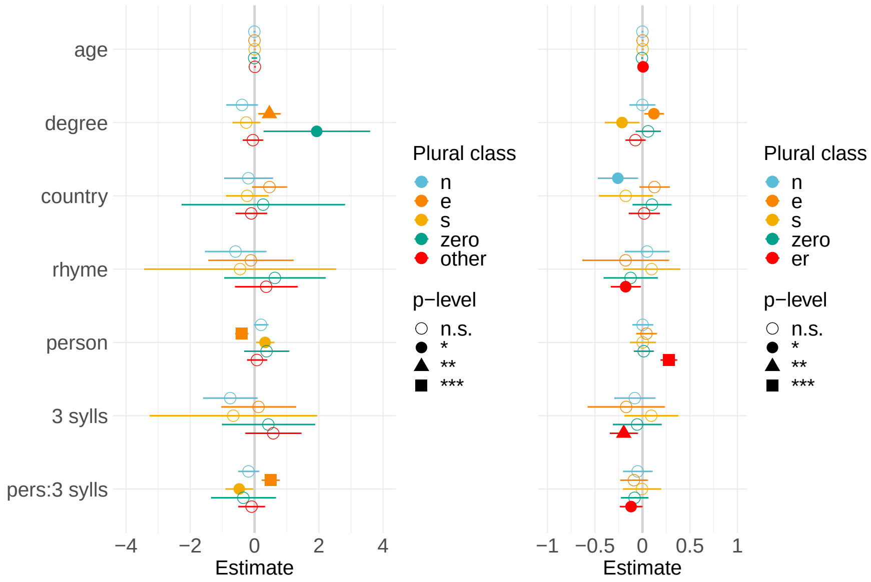


Figure 3.1: Visualisation of estimated coefficients in GLMM for production probabilities (left panel) and LMM for z-scaled ratings (right panel). *p<0.05; **p<0.01; ***p<0.001

lowing Zaretsky and Lange (2015); McCurdy (2019)) as implemented in the R (R Core Team, 2020) package *lme4* (Bates et al., 2015). A binomial model was run for each plural class with as dependent variable the plural class. We collapsed *uml* and non-*uml* classes since the *uml-er* and *uml-∅* classes had too few positive data points to converge (also improving comparability to McCurdy (2019) who did the same because of confounds). Independent, fixed variables were semantics (person/object), number of syllables (two/three), the interaction of these factors, rhyme (rhyme/no rhyme), university degree (yes/no), age and finally country of primary education (Germany/Elsewhere)[9]. Participant and word were included as random effects. All fixed effects are visualised in the left panel of Figure 3.1, and the full model can be found in Appendix C. Note that the model for *-er* plurals did not have enough data points to converge and was therefore excluded from the analysis.

The statistical model showed a positive effect of holding a university degree on preferring *-e* plurals. No effect of age was found. Rhyme and country did not have an effect on any of the plural classes (see also Figure 3.2, left panel). Even though there are (slight) differences in preferences for different plural classes in two- vs three-syllable words (see Figure 3.2, center panel) the model does not show a clear effect (Figure 3.2, right panel). There is no reliable influence of a word describing a person on the use of *-n* plurals ($p = 0.07$), but a positive one *-s* plurals ($p = 0.02$) and a clear

[9] Country had to be Germany/Elsewhere because of data sparsity.

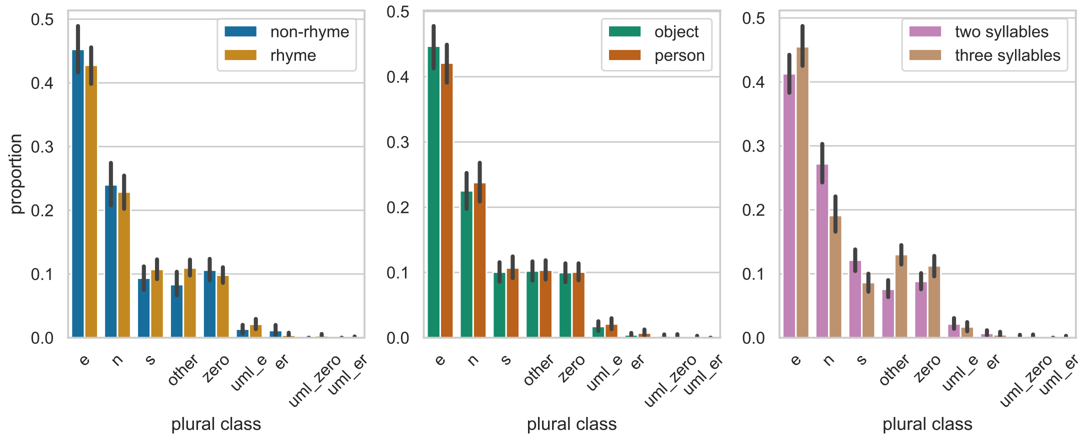


Figure 3.2: Influence of rhyme, semantics and number of syllables on production of plural classes in 197 German speakers. Error bars indicate 95% confidence intervall across speakers.

negative one on -*e* plurals ($p < 0.001$), with the opposite direction for the interaction of describing person and three syllables.

We investigated whether the effects of semantics on -*e* and -*s* plurals are reflected in the corpus data, using the same method as described in Section 3.1. The regression models indeed identified a negative effect of the *person* semantic class on the -*e* ($p < 0.001$), but also a negative, yet weaker one on the -*s* class ($p < 0.001$) (effects visualised in Figure 3.3, models in Appendix C).

No priming effect of surrounding noun classes could be found. Including whether the last noun in the description sentences had the same plural class as produced by the participants ("last plural class") prevented the regression models from converging. We therefore treated it as an alternative to the semantic and syllable factors and excluded those. There was no statistically reliable effect of *last plural class* on any of the endings (full models in Appendix C).

3.3.2 Rating task

Overall. Participants' ratings deviated from their productions and from previous work. They showed an overall preference for the -*e* class (M 3.7), followed by -*n* (M 3.6) and -$\varnothing$ (M 3.2) (visualised in Figure 3.4), thus deviating from the typing task where -*s* was preferred over -$\varnothing$. This also differs clearly from the results of previous work. While in McCurdy (2019) the -*n* ending was favoured (M 3.8), the participants of Zaretsky and Lange (2015) preferred -*s* overall (M 3.7). When dividing the data into person

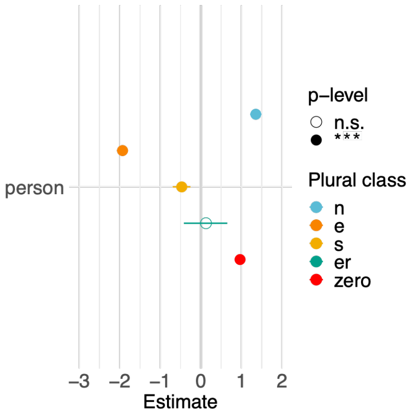


Figure 3.3: Visualisation of GLMM for the effect of *person* on the use of plural classes in the 'semantic dataset'.

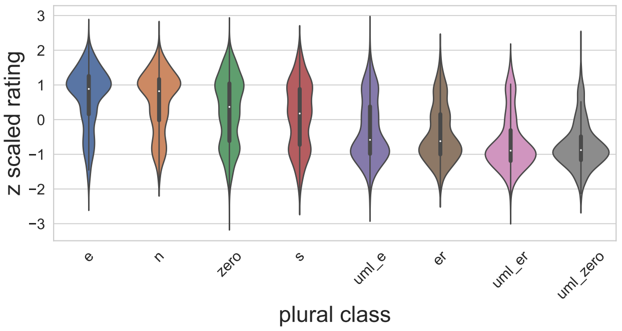


Figure 3.4: Distribution of z-scaled ratings by 197 German speakers across plural classes. The wider the 'violin', the more participants rated respectively.

and non-person, and two- and three-syllable words, and when excluding the two faulty items *Walgimirz* and *Zackabat*, the same trend emerged.

Statistical model. In order to again assess the influence of the various demographic factors and independent variables we made use of the same method as McCurdy (2019): we ran a Linear Mixed Model (LMM) from the *lme4* package, and ratings were z-scaled for each participant, in order to make ratings comparable and real-valued. *uml* and non-*uml* were collapsed to make results comparable to the productions. These ratings where then used as dependent variables, with a separate model for each plural class. Independent, fixed variables and random effects were the same as in the binomial models for the production task. As can be seen in the right panel of Figure 3.1, there is a tendency that participants holding a university degree disfavour *-s* ($p = 0.02$) and favour *-e* plurals ($p = 0.02$). There is also a very small effect of age on the use of *-er* plurals ($p = 0.03$). Participants educated in Germany show lower ratings for *-n* plurals ($p = 0.02$). The reason for this was not investigated further, since it was not a focus of the present study. There is no clear effect of rhyme, apart from a slight negative effect on *-er* plurals ($p = 0.04$). Again, we do not see the expected effect of semantics or the interaction of semantics and number of syllables on the rating of *-n* plurals ($p = 0.98$). The only effects we do see is a lower rating of *-er* plurals in three-syllable words ($p = 0.01$) and, notably, a clearly higher rating of *-er* plurals in person words ($p < 0.001$). Without the two faulty items (*Walgimirz, Zackabat*) all effects except the interaction between semantics and number of syllables on *-er* plurals were retained.

3.4 Discussion

The distribution of production classes is similar to previous studies. This indicates that the observed pattern is stable across various selections of wug words. In the rating task, however, responses differ from McCurdy (2019); Zaretsky and Lange (2015); the problems of rating tasks to explore pluralisation strategies will be discussed below. The distribution of plural classes in production cannot be explained by the distribution of plurals in corpus data. However, participants might either match token instead of type frequency, or the probability of plural endings in rare words (at the tail of the frequency distribution), instead of the entire distribution. This was not investigated further.

The study could not confirm McCurdy (2019)'s influence of age on *-s* or -∅ endings. Notably, McCurdy (2019)'s *English knowledge* factor has the same effect as the *degree* factor in the present study, suggesting a confound. Dabrowska (2008) showed that education and size of vocabulary can have a considerable effect on speakers' ability to provide inflected forms of wug words in Polish. While German participants did not have any difficulties in providing plural forms per se, the influence of education seems to manifest itself in a (dis-)preference for certain plural classes.

The effects of rhyme deviate from McCurdy (2019). The only effect in the present study is a negative one on *-er* plurals in both rating and production task - contrary to McCurdy (2019) where the effect was positive. Moreover, there is no effect on *-s* plurals, in McCurdy (2019) it had a clear negative one in both tasks. Marcus et al. (1995) had predicted a negative effect of rhyme on the use of *-s* endings, assuming that the more unfamiliar non-rhyme words would trigger a default response. This prediction is not supported by our data, providing more evidence against their claim.

Notably, neither the rating nor the production data support the main hypothesis of this study, the influence of semantic category on the production of *-n* plurals. On the other hand, there is clear negative effect of semantic category on *-e* and a smaller and less reliable positive effect on *-s* plurals. Further exploration of the corpus data revealed that a similar effect of *person* on *-e* plurals can be found there, but the effect on *-s* points into the opposite direction. Speakers' responses thus corresponded only partly to semantic influences in corpus data. Moreover, these effects are not at all visible in the rating data. Possible reasons for these phenomena might be a) that the underlying observations in the corpus data only hold for type but not token frequency and participants might match the latter. Perhaps b) the effect of semantics is only visible in the speaker data if it is sufficiently strong in the corpus data (as the effect on *-e* is somewhat stronger than on *-n* , see Figure 3.3). However, this would not explain the lack of effect on ratings. It is c) furthermore possible that participants do not take into account semantics at all when producing plural forms and the effects of semantics on *-e* productions and on *-er* ratings are due to confounds with factors not controlled in this study. A final explanation might be d) that semantic information is only one among many factors which influence the selection of plural class and that possibly the effect of person/object semantics only comes into play in specific circumstances, as it apparently has for *-e* plurals, but not in others (*-s* and *-n*), such as an interaction between semantics and phonology not captured by our wug words.

The higher ratings of *-er* in the *person* category suggest both an influence of semantics, and a possible inadequacy of rating tasks to examine pluralisation strategies. In the German Wiktionary dataset (described in Sections 3.1 and 4.2) plural forms ending in *-er* mostly belong to the *-∅* class (78%) where the singular ends in *-er* as well, not the *-er* plural class (*uml-er* 15%, *-er* 6%). The semantics most common in these words is *person* (58%). *-er* is commonly used in German to derive a term describing a person from an activity or noun ("Nomen Agentis", e.g. *laufen (to run) → der Läufer (runner)*; Baeskow (2011)). It is thus not surprising that speakers strongly associate *-er* with the semantic category of person - after all, these co-occur very frequently - even though they usually also require an *-er* singular which none of our wug words has. Thus, the high ratings for *-er* in the person condition can only be explained if one assumes that participants ignore or forget the singular form of the word while rating its plural form. While this does suggest an influence of semantics, it also calls into question the validity of rating tasks for measuring how Germans inflect words, a method already criticised by Zaretsky and Lange (2015, p. 5) as exploring "the whole spectrum of their associations or creativity" rather than "their internalized pluralization strategies". This might also be an explanation for the deviation between productions and ratings observed in this study and previous work (McCurdy, 2019).

The overall effect of both broader semantics and more specifically person/object categories is therefore somewhat unclear. The interaction between semantics and plural classes found in corpus data is only partly put to use in plural productions and ratings of wug words. This suggests that the present study has only partly captured the influences at play in German pluralisation, and that the precise factors governing plural productions remain unclear.

Chapter 4

Modelling German inflection with Encoder-Decoder models

One of the most recent architectures used to model inflection are Encoder-Decoder models (Sutskever et al., 2014), initially for the English past tense (Kirov and Cotterell, 2018), and recently also for the German plural system (McCurdy, 2019; McCurdy et al., 2020). However, the results were mixed (Corkery et al., 2019; McCurdy, 2019). While McCurdy (2019) found a broad correlation between production probabilities of model and speaker data, the model showed some clear failure modes, such as overestimating the prevalence of the *-e* class and underestimating the *-n* class. Additionally, the model showed considerably lower variability both across and within models and items than speaker data.

We hypothesised that the model's ability to predict speaker data can be improved by adding additional information available to the speakers: the meaning of the words. This chapter will give an overview of the model architecture and techniques used to add semantic information. Subsequently, the model's predictions both on held-out test data as well as the wug word data set used in the speaker experiment will be presented and analysed.

4.1 Enhancing an Encoder-Decoder model with semantic information

Encoder-Decoder models consist of two parts: In the encoder part a Recurrent Neural Network (RNN) (Rumelhart et al., 1986) reads in the target word in its singular form, with an RNN in the decoder part producing the target plural form letter by letter. The RNNs are usually implemented using so-called Long-Short Term Memory cells (LSTMs) in order to avoid vanishing gradients (Hochreiter and Schmidhuber, 1997). Additionally, an attention mechanism is often included, designed to enable the decoder to pay attention to specific parts of the input (Bahdanau et al., 2015). Because the most likely letter at each position does not necessarily lead to the most probable word, a beam search decoder is used to decode the most likely final output word.

The observation by Gaeta (2008) includes grammatical gender. Additionally, this feature is usually available to speakers and is informative of plural class (see Section 2.1.1). We thus included grammatical gender into our model but - contrary to McCurdy et al. (2020) - not as ‘prefix token’ before the singular word form, but rather utilised a model architecture with parallel inputs, with one of them being the word and a second input the word’s gender. This means that the gender is available to the model at each time step and does not have to be memorised over the entire input.

In order to test the influence of semantics, we also needed to add semantic information. Various methods have been used in the literature to represent semantics in computational models. For example, in Joanisse and Seidenberg (1999)’s model of English inflection each word was simply represented as a different one-hot vector. This approach cannot express similarities between words and is thus inappropriate here. Another technique would be to use word embeddings (e.g. *word2vec* (Mikolov et al., 2013)), as in Carter et al. (2019); Williams et al. (2020). They might, however, include hidden morphological information and make the attribution to semantics uncertain. Using semantic representations compiled by using human similarity judgements such as McRae features (McRae et al., 2005) would be more psychologically plausible, but they currently cover not enough words to be usable to train a Neural Network. We therefore settled on using image embeddings. These have been found to be significantly correlated with human similarity judgements (Peterson et al., 2016) and have been shown to be useful in other multimodal tasks (e.g. Gella et al. (2016), also in cognitive modelling Gella and Keller (2018)).

Our second method to represent semantics made use of the noun hierarchy WordNet (Fellbaum, 1998). Here, the meaning of a word is a category it belongs to within WordNet. For example, a *cup* might belong to the category *crockery*, or when further ascending the hierarchy, *ware* and eventually *object*. While this method cannot represent the fine-grained similarities between meanings as embeddings do, the granularity of representation can be chosen freely.

The representation of semantics was used as a further input to the model, thus again providing the semantics at every time step during encoding.

4.2 Data

Words and their plural forms were taken from a corpus scraped from Wiktionary[1]which includes gender information. We created two datasets:
Image Dataset: Each German noun was translated to English if possible (by using a German-English dictionary scraped from the English Wiktionary[2]). If the translation of the noun was included in ImageNet, we tried to download images from the first 100 links provided by ImageNet. If every single one failed, we tried the next 100 links. Images from the ESP game dataset (Von Ahn and Dabbish, 2005) were added. This resulted in 389,953 images overall (ranging from 1 to 764 per word, M 60.35, SD 46.62), but reduced the number of distinct German nouns to 6,511. In order to create image embeddings, we extracted the activations of the last (average pool) layer (as in Peterson et al., 2016) from a ResNet-18 (He et al., 2016), trained on ImageNet (Deng et al., 2009) and part of the PyTorch library (Paszke et al., 2019)[3], resulting in embeddings of size 512. For testing the image embeddings we extracted semantic category information. We searched WordNet (Fellbaum, 1998) for the English translations and assigned them to one of six classes according to their hypernyms (*object 53%, person 25%, animal 11%, matter 7%, attribute 4%, other 2%*). The classes were chosen such that they include the classes relevant for the hypothesis and then trying to cover further classes to reduce the size of the *other* class.
Semantic dataset: We again used all German nouns and added their semantic information from WordNet (thus avoiding the limitations of ImageNet) (*object 22%, person 21%, attribute 11%, matter 7%, animal 4%, other 35%*), resulting in a dataset with 15,383 distinct German nouns.
Nouns were presented in their orthographic form, as German's phonology largely follows its orthography. Both the image and the semantic dataset were divided into 10% test and 90% training data, of which 10% where used as validation data.

[1] https://github.com/gambolputty/german_nouns

[2] https://github.com/karoly-varasdi/de-wiktionary-parser; translating the words via this method was necessary since WordNet does not, as yet, include a German version.

[3] https://pytorch.org/docs/0.2.0/_modules/torchvision/models/resnet.html

4.3 Testing image embeddings

Our model presupposes that the image embeddings include sufficient semantic information. To test this, we used a dense feedforward network, implemented in Keras (Chollet et al., 2015). Inputs were the embeddings and targets their associated word from the corpus. As data we used the training data from the picture dataset, with 3205 target classes, and an 80/10/10 train/val/test split[4]. We tested one, two and three hidden layers with hidden dimensions of 20, 100 and 512 and found an architecture with one hidden layer with 512 hidden dimensions to work the best. Such a model achieved a test accuracy of 24.2% (majority baseline 0.1%). We assumed this rather poor performance was mostly due to the fine-grained distinction between classes that the image embeddings were not created for (the ImageNet data set they were trained on only included 1000 classes). Next, we therefore made the categories more coarse by ascending the hypernym hierarchy in WordNet[5]. Reduced to 923 classes, test accuracy was 41.1% (majority baseline 14.5%). When just predicting the six classes described above, a test accuracy of 82.0% was achieved (majority baseline 53.2%).

The mistakes of the model were semantically intuitive. In order to avoid mistakes due to overfitting to a majority class, we subsampled the 923-class data set such that each class included 1000 images (17 classes overall remaining). The reduced training data set resulted in a comparatively low test accuracy (59.3% only testing on classes present in the training set, majority baseline 25.2%). The highest probability of confusion relative to its true label had *ingredient* with *dish* (18%), followed by *device* with *implement* (15%) and vice versa (14%) as well as *equipment* with *device* (15%). *Vegetable* tended to be confused with *plant* and *plant_organ* (14% each). Most mistakes tended therefore to broadly be within semantic classes. The information included in the image embeddings is thus presumably sufficient to cover the person/object distinction necessary to test our hypothesis.

[4] In the main models we are ultimatively interested in the phonology, thus there is overlap between images in the training and test/val data, and therefore the original validation data cannot be used here.

[5] Used was the hypernym 6 levels down from the uppermost.

4.4 Wug tests on ED models

4.4.1 Methods

4.4.1.1 Hyperparameters

While following the general model architecture of Kirov and Cotterell (2018); McCurdy (2019), we slightly changed the model in order to increase training efficiency. The input sizes were 64 for phonology and 20 for gender. The third input varied depending on the model architecture we explored:

Architecture 1: An architecture with image embeddings as semantic representation. The image embeddings of size 512 were fed through a linear layer first to reduce them to a length of 20.

Architecture 2: An architecture with semantic categories as semantic representation. The size of the semantic input was 20.

Architecture 3: An architecture without semantic information. Here no third input was included.

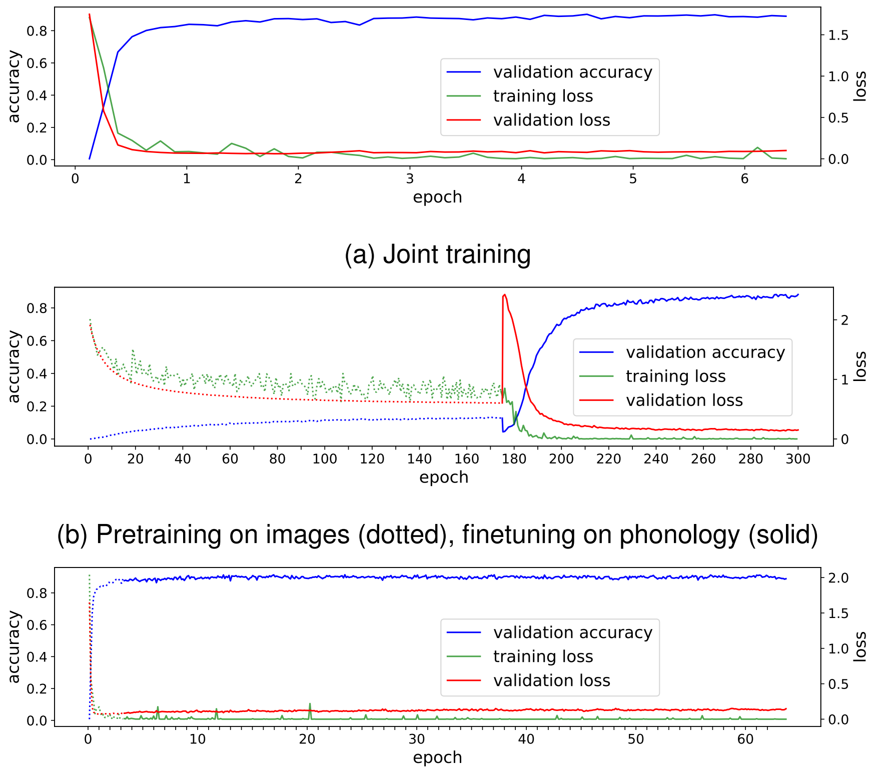
Both encoder and decoder had 2 layers with a hidden size of 100 each, and a between-layer dropout of 0.3. Both RNNs were implemented as LSTMs and were unidirectional following McCurdy (2019) (and contrary to Kirov and Cotterell (2018)) since their's were the results we were most interested in comparing our data to. The decoder used the global attention mechanism by Luong et al. (2015). We used Adam with a learning rate of 0.0002. For decoding we used beam search with a beam size of 12.

4.4.1.2 Training a multimodal model

In the following we will discuss various methods we experimented with to achieve best possible performance on a multimodal model where we attempted to use image embeddings as semantic representation (architecture 1).

Joint training of all inputs. Initial experiments showed that the model started to overfit to the training data after only two epochs, presumably because a word occurs on average 60 times within an epoch (see training loss in Figure 4.1a). After convergence on the validation data at 6.4 epochs, test accuracy was 87.4%. We took this as evidence that the model started to ignore the images as the less informative source of information.

Pretraining on image embeddings. Phonology and gender were masked with



(a) Joint training

(b) Pretraining on images (dotted), finetuning on phonology (solid)

(c) Pretraining on phonology (dotted), finetuning on image embeddings (solid)

Figure 4.1: Training curves for various training regimes of a multimodal model

x during the first half of training.[6] After pretraining, a test accuracy of 13.6% was reached, having converged on the validation data at 175 epochs. While the model struggled with fine semantic differences, errors were often within semantic classes (e.g. predicting *house mouse* instead of *field mouse* or *olive* instead of *pear*), but frequently also totally unrelated (e.g. *swimmer* instead of *glove*). This might be due to the difficulties in differentiating between fine semantic distinctions already discussed in Section 4.3. Further training on phonology increased the accuracy to 87.4% after 300 epochs (see Figure 4.1b). While this approach was promising, training a single model this way took about 36h on a single GPU. Moreover, the weights of the first layer suggested the model tended to utilise the semantic information less and less, the more it was trained on phonology.

Pretraining on phonology. We pretrained on phonology (masking the image em-

[6] Since the validation and test data of the Image dataset necessarily are made-up of words not in the training set, we used data from the training set and split it into training, validation and test set respectively (.81/.09/.1) for pretraining on images. This also results in the sudden spikes in loss and accuracy when switching to phonology in Figure 4.1b.

beddings) and after convergence (3.2 full epochs, but words appear more than once within one epoch, thus showing each word on average 190 times) added the image embeddings to the training (see Figure 4.1c). While this improved the training speed of the model, the improvement to the validation accuracy was small (88.2% vs 86.9% after pretraining) and was presumably due to the pretrained model being stopped too early, or else random variation, since a model only trained on phonology for an equal number of epochs got a similar test accuracy (88.3%). We proceeded to test this model on the wug data and found no evidence of any influence of semantics (all words received the same ending no matter whether they were presented with person or with object images). To ensure that this was due to semantics not being informative as opposed to the model not being able to learn from the image embeddings, we ran a further model with the semantic dataset as training data. The model was jointly trained on phonology and semantics. Here, the effect of semantics was clearly visible (see below in Section 4.4.2). We thus concluded the following: while there is enough semantic information in the image embeddings to detect the influence of semantics in theory (see Section 4.3), the model apparently was unable to learn this. Possible explanations are that there is not enough training data or that the way the image embeddings were introduced (as parallel inputs to the encoder) was not ideal for the model to learn. Alternatives, such as adding the image embeddings only to the decoder or to the softmax layer were beyond the scope of this project (a comparison between different approaches can be found for example in Tanti et al. (2017)).

For our final analysis, we therefore decided to use the semantic dataset, and thus model architecture 2. This has multiple advantages: a) the dataset includes more words (see Section 4.2), b) training is significantly faster (≈40 mins vs ≈13h), thus enabling us to increase the number of random simulations we were able to run and finally c) the model showed an influence of semantics we could actually compare to the speaker experiment data.

4.4.1.3 Final training regime

The models were trained to convergence on the validation set at 100 epochs, which corresponded to a training accuracy of about 97%[7]. As in McCurdy (2019), we ran mul-

[7] This does not fully correspond to McCurdy (2019); Kirov and Cotterell (2018) who trained until a training accuracy of more than 99% was reached, arguing that humans have completely memorised the training data. However, we trained 10 models for an additional 100 epochs (training accuracy M 98.5%, SD 0.006) and found the test accuracy (M 88.2%, SD 0.01) to be slightly worse compared to the test accuracy of ten models trained to only 100 epochs (M 89.9, SD 0.004). Moreover, we found the effects

tiple simulations, initialised with different random seeds since Corkery et al. (2019); McCurdy (2019) showed that the behaviour of Encoder-Decoder models on wug data is severely influenced by initialisation. Because our experiments had two conditions, with 100/99 participants in each, we ran 100 simulations initialised with different random seeds and tested the models on both conditions (thus testing each wug word once as person and once as object) in order to save on training time.

For reference, we trained an additional 10 models only on phonology and grammatical gender ('reference models', architecture 3). The models were trained to convergence at 100 epochs, the same as in the other models, and their average training accuracy was 96% (SD 0.007).

4.4.2 Results and discussion

The main interest of this thesis lies in the fit of the model to the speaker data and will be discussed in Chapter 5. We will here only give a brief overview and discussion over key findings from the model predictions on both held-out test data and wug data.

4.4.2.1 Results on held-out test data

McCurdy (2019) already discussed the general performance on held-out test data in ED models trained on German plurals in detail. We therefore only briefly describe some main observations in 10 ED models, with an emphasis on the influence of semantics.

Accuracy and errors. The models' mean test accuracy was very similar to previous work and that of the reference models, indicating that semantics did not have a positive effect on held-out test performance. The models reached an average test accuracy of 89.9% (SD 0.004[8]), which is slightly better than McCurdy (2019) (87.3%) and McCurdy et al. (2020) (88.8%). This is also very similar to the average accuracy of the reference models trained only on phonology and grammatical gender (89.2%, SD 0.007).

Influence of semantics. Semantics did not show a clear pattern of influence on accuracy or errors. Test accuracy varied across the different semantic classes but without a discernible pattern (*person* M 91.0%, SD 0.007; *object* M 88.5%, SD 0.012; *matter*

discussed in Section 4.4.2 to be very similar. Since we operated under time and resource constraints we thus opted to rather run a higher number of different simulations. We nonetheless discuss possible effects of this decision in Chapter 5 and the entire statistical analysis for the 10 fully trained models can be found in Appendix C.

[8] If not indicated otherwise, standard deviations are reported across models.

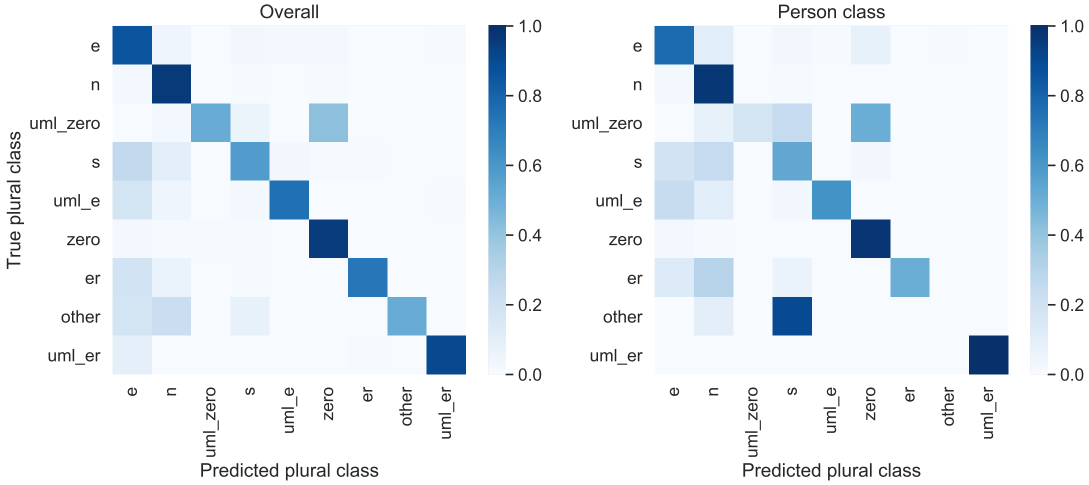


Figure 4.2: Confusion matrix of true versus predicted plural class in held-out test data, for all words (left) and for *person* words (right). The darker the colour, the higher the percentage of words from the *true class* predicted as the *predicted class*. For perfect accuracy, only the diagonal would be coloured.

M 85.2%, SD 0.02; *attribute* M 93.7%, SD 0.016; *animal* M 87.3%, SD 0.022; *other* M 90.0%, SD 0.008). As can be seen in the left panel of Figure 4.2, the models often overgeneralise to the -*e* and -*n* class and omit *uml*, especially in the -∅ case. This is the same in the *person* class, albeit with a slightly higher tendency to generalise to -*n* (right panel of Figure 4.2). This is expected given that *person* is highly predictive for -*n* plurals (see Section 3.1). However, the *other* class is also much more likely to be mistaken with the -*s* class (90%) among *person* words than among all words taken together (8%). This could be interpreted as evidence in favour of -*s* as a default class, but the lack of this behaviour generally calls this into question.

4.4.2.2 Results on wug data

In general, the models favoured the -*e* plural class (M 47%, SD 0.08), followed by -*n* (M 16%, SD 0.08) and -*s* (M 16%, 0.05). As indicated by the SDs already, variability across models was relatively high. A close comparison between plural distributions in model data and the data in the speaker experiment as well as a discussion of variability can be found in Chapter 5. We will here only note general observations applicable to the model without comparison to the speaker data, focusing on the influences of rhyme, semantics, and number of syllables (an overview over the data can be found in Figure 4.3). In order to test these effects in the productions we ran

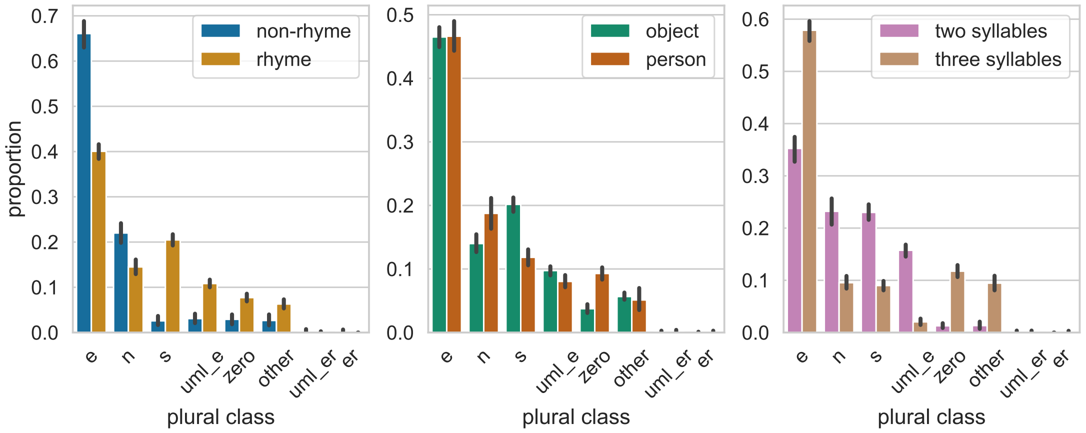


Figure 4.3: Influence of rhyme, semantics and number of syllables on distribution of plural classes in 100 ED models (averaged means across models). Error bars indicate 95% confidence interval across models.

the same statistical model as in Chapter 3 for the speaker data: a Generalised Linear Mixed Model (GLMM) for each plural class (again collapsing *uml* and no *uml*) with rhyme (rhyme/no rhyme), semantics (person/object), number of syllables (two/three) and their interaction as predictors. Model seed and word were included as random effects. Model coefficients are visualised in Figure 4.4. The model's entire statistics can be found in Appendix C.

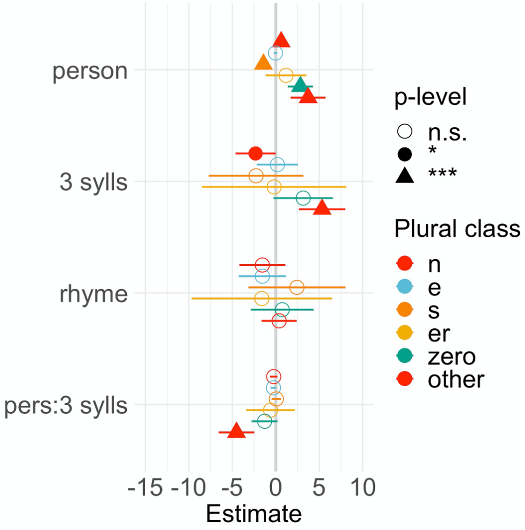


Figure 4.4: Visualisation of estimated coefficients in GLMM for production probabilities in ED models. *p<0.05;***p<0.001

Influence of rhyme. There was no statistically significant effect of rhyme on any of the classes. Even though the differences between rhyme and non-rhyme seem to be quite high within the individual classes (e.g. *-e* 40 vs. 66%, *-s* 20 vs 3%, see Figure 4.3), the statistical models do not confirm this. The large difference seems to be caused by word-specific rather than systematic effects, for example, the reluctance of models to use *-s* in the non-rhyme condition is largely driven by individual items ending in *-s* or *-z* (see also Figure 5.2a). Our model does therefore clearly not use *-s* as a default plural class, contrary to the hypothesis by Marcus et al. (1995).

Influence of semantics and number of syllables. Semantic class has the expected effect on the use of *-n* endings, confirming the observation by Gaeta (2008). Whether a word is presented as *person* influences the production of *-n* ($p < 0.001$), -∅ ($p < 0.001$) and *other* ($p < 0.001$) endings positively and of *-s* plurals negatively ($p < 0.001$). A possible explanation for the unexpected effect of -∅ is its prevalence with *person* words (see Section 3.4). A positive effect of *person* on the use of -∅ plurals is also present in the corpus data (Figure 3.3, $p < 0.001$). The proposed interaction between number of syllables and semantics is not supported by the data, and neither is an effect of number of syllables by itself. These only seem to have an effect on the *other* class (both $p < 0.001$), whose diversity makes it hard to give clear reasons for this. A possible reason for the lack of influence on *-n* plurals is that Gaeta (2008)'s hypothesis actually applies to the *weak* inflection class as opposed to *-n* plurals in general.

4.4.2.3 Summary

We identified two main trends: Firstly, no clear effect of semantics on the performance of existing words was discernible. Secondly, there was no statistically significant effect of either rhyme or number of syllables on the wug data, but there was a clear positive effect of the semantic *person* class on the *-n* and -∅ and a negative effect on the *-s* class.

Chapter 5

Joint Analysis and Overall Discussion

In the previous chapters we have largely focused on exploring effects on the German speakers and ED models individually. Ultimately though, the goal of cognitive modelling is to predict and explain human behaviour. The model should thus be measured in how much its behaviour resembles that of humans. In this chapter we will therefore first analyse differences and similarities between model and speaker data. Subsequently, we will discuss the main findings from the previous chapters as well as the joint analysis to find answers to our second question: does including semantics into a model of inflection help to predict human behaviour?

5.1 Joint Analysis

There are various ways in which to analyse and evaluate the model to speaker data fit. Our analysis here will focus on four main points: a) Do the models show similar behaviour and failures as in previous work? b) How much do model and speaker production probabilities correlate, compared to previous work and models without semantics? c) Does the variability in the models resemble the variability in the speaker data? And finally d) what is the effect of training differences between the current study and McCurdy (2019)? All of these points will be examined with a special focus on the influence of semantics.

5.1.1 General behaviour compared to previous work

Aggregating the respective datasets, the models as well as the speakers have a clear preference for the *-e* plural class overall (47% vs 43%), followed by *-n* plurals (16%

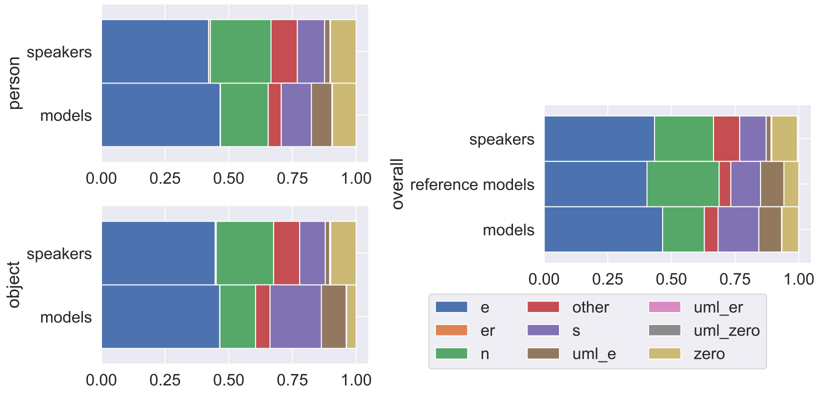


Figure 5.1: Comparing production probabilities across models and semantic categories. 'reference models' are trained without semantic information.

vs 23%). The model ranks -*s* plurals next (16%) whereas in the speaker data the large *other* class follows (11%) which is then followed by -*s* (10%) (see Figure 5.1, right panel, speakers vs models). It is instructive to compare the results from our speaker and model experiments to McCurdy (2019) to work out general tendencies of ED models and German speakers across different stimuli sets.

Our results do not show the same failure modes as in McCurdy (2019), such as a lack of -*n* endings in the model. We cannot directly compare the distributions to McCurdy (2019) because only 25% of words in our data rhyme as opposed to 50% in McCurdy (2019)'s and they found a large effect of rhyme on their model production probabilities. We thus only consider distributions within rhyme condition in the following. While McCurdy (2019)'s models use too many -*e* plurals in both rhyming conditions, our models only overestimate their number in the non-rhyme condition (non-rhyme: models 66%, speakers 44%; rhyme: models 40%, speakers 42%). Note that this cannot be explained by additional overfitting to the training data in McCurdy (2019), since the 10 fully trained models also produced -*e* in 40% (67%) of all rhyme (non-rhyme) cases. McCurdy (2019) reported a difference in the -*n* plurals across rhyme condition (rhyme 25%, non-rhyme 35%), while there was no difference in the ED model data (both 15%). In our data, however, participants show very little difference (23 vs 24%) but the model does (15 vs 22%).

5.1.2 Correlation between model and speaker data

To explore how the models' predictions correlate with the speaker data, Albright and Hayes (2003); Kirov and Cotterell (2018); Corkery et al. (2019); McCurdy (2019) used

plural class	overall	reference
n	0.65***	0.58**
e	0.49*	0.56**
uml-e	0.39	0.47
other	0.53**	0.42*
zero	0.42*	0.4
s	0.78***	0.68**
er	-0.24	nan
uml-er	-1.0***	nan
uml-zero	nan	nan

Table 5.1: Spearman rank correlation between model and speaker data, for the main models ('overall') and for the reference models trained without semantic information ('reference'). 'nan' indicates that the plural class was missing in either the model or speaker data. Note that this table includes results from 10 reference vs. 100 main models, and is thus not necessarily comparable. $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

a spearman rank correlation between the production probabilities of the plural classes in speaker and model data. The correlation between our model and speaker data was slightly higher than in previous work. We found a correlation of .72, compared to .62 in McCurdy (2019). In the spirit of Albright and Hayes (2003) who divided the data into regular and irregular classes for this purpose, we followed Kirov and Cotterell (2018); McCurdy (2019) and also calculated correlation within each plural class. The results can be seen in Table 5.1 ('overall'). Results within classes are mostly broadly correlated. The negative correlations of *-er* and *uml-er* are due to their infrequency in the model data (3 occurrences). The question remains whether the overall improvement in correlation can be attributed to the inclusion of semantics in the model.

Correlation with the main models was slightly higher than with the reference models which were trained on phonology and grammatical gender only. The overall spearman correlation between reference models and the speaker data is .65, thus slightly lower than for the models including semantics. Again breaking down by plural class, the semantic models show higher correlation for the *-n* , *-s* , -∅ and *other* classes, and the other way around for *-e* and *uml-e* (see Table 5.1, 'reference'). The higher number of missing classes might be a clue that the reference models produce lower variety of different plural classes (see below). We can thus tentatively conclude that the semantic models are better correlated with the speaker data - however, to draw final conclusions, a higher number of reference models would have to be trained.

Possibly, semantics have a beneficial effect as they avoid some of the tendencies of the reference models. The right panel of Figure 5.1 shows the distribution of endings both across models, speakers and reference models. The poorer fit of the reference models seems to be due to a rather strong overestimation of *-n* and *uml-e* endings and underestimation of -∅ plurals.

5.1.3 Data variability

Production probabilities can vary in multiple ways. On the speaker/model level, answers **across** participants/seeds differ if speakers/models show very different production behaviour from each other. The more answers vary **within** speakers/models, the more different classes a single participant/seed produces. The same distinction can be made on the word level. If there is a big difference **across** words, plural class distributions are not the same for every word. If variability is high **within** words, speakers/models produce a high number of different plural classes for a word.

McCurdy (2019) reported that the variability both within words and across models was considerably lower than in the speaker data. This might be interpreted as a failure of the models - seemingly not capturing a key characteristic of the speaker data. We therefore analysed whether this was also the case when including semantic information.

5.1.3.1 Variability across and within words

As in McCurdy (2019), the variability both within and across items was high in both model and speaker data, though generally higher in the speaker data (see Figure 5.2a). On the one hand, variability across words was high, as e.g. the *-e* class was generally dominant, but some words show a clear preference for other endings (e.g. *Reslans, Jakaselb, Femmotak*). The number of different classes per word ranged from 2 (*Pobekus*, *Filast*) to 6 (*Femmotak*, *Karagul*) (M 3.9, SD 1.1) in the model, and the lowest number was 4 (*Walgimirz*) and ranged up to 8 (*Femmotak*, *Toldar*) (M 6.3, SD 0.9) in the speaker data. This also indicates the higher variability within words in the speaker than in the model data. Interestingly, in both data sets the number of different plural classes per word was on average higher in the *person* condition than in the *object* condition (model: object M 2.8, person M 3.8; speakers: object 5.7, person 5.9).

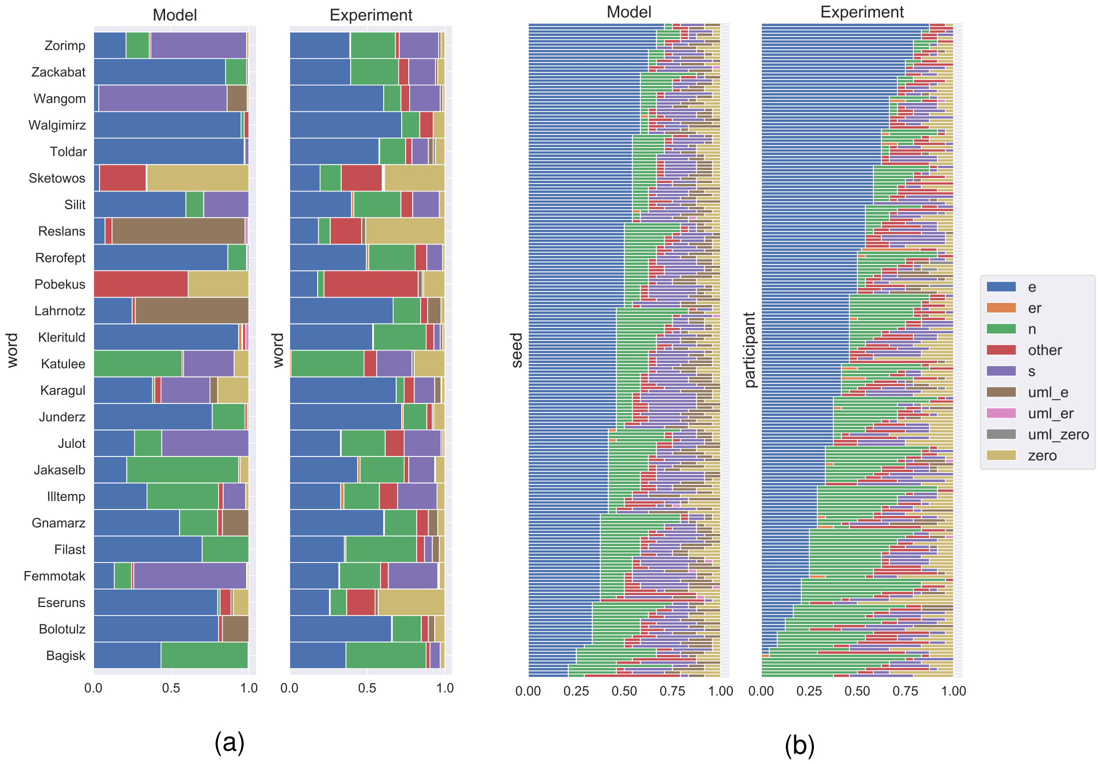


Figure 5.2: Use of plural classes in 100 models and 197 German speakers. Variability is higher in the latter, both across and within words (a) and speakers/models (b). Note that since models were tested in both conditions, each appears twice in (b).

The variability within words in the model data seems to be generally higher than in McCurdy (2019). In a subset of 50 of their ED models, the mean number of different plural classes per word was 3.0 (SD 0.83). In the 25 ED models (which include information about grammatical gender) in McCurdy et al. (2020) that number rose to 3.2 (SD 1.3). Since the results might vary because of the use of different wug words, we compared to the reference models. The number of different plural classes per words is clearly lower (M 2.2, SD 0.48). These results suggest an overall trend of fewer plural class cues leading to lower variability. Since the number of reference models was quite low, replication is necessary to draw final conclusions.

5.1.3.2 Model and speaker variability

As in McCurdy (2019), speaker variability was considerably higher than model variability. As can be seen in Figure 5.2b, the models generally did not reach the extremes of the speakers' distributions. While for example the use of *-e* in the models reached from 25% to 63%, it ranged all the way from 0 to 88% in the speakers. *-n* was used in

2 to 38% of cases in the model, but from 0 to 79% by speakers. On the other hand, the range of *-s* was quite similar (model: 4 to 31%, speakers: 0 to 38%).

As with within-word variability, within-model variability was also higher in the present study than in McCurdy (2019); McCurdy et al. (2020). Mean number of different plural classes per model was 6.0 (SD 0.41) in the present study, but only 4.8 (SD 0.97) and 5.0 (SD 0.76) in McCurdy (2019) and McCurdy et al. (2020) respectively[1]. However, the reference models also showed quite high variability within models (M 5.9, SD 0.42). The within-model variability is thus more likely due to the choice of wug words rather than any influence of semantics.

5.1.4 Effect of training differences

As noted in Section 4.4.1.3, our training regime differed from McCurdy (2019); McCurdy et al. (2020) in that we do not train to full convergence on the training set. Corkery et al. (2019) noted that the correlation between model and speaker data was generally lower when the ED models were trained to full convergence on the training data. This begs the question whether our findings might be a result of the different training regime rather than of other effects, such as semantics. And indeed, correlation between the 10 fully trained models and speaker data is lower (.67)[2]. While the within-word variability decreases quite dramatically in the fully trained models (M 2.7, SD 1.1) which might call into question the observation that variability increases with the inclusion of additional cues, the within-model variability does so only slightly (M 5.9, SD 0.74), confirming the overall trend of higher variability the more potential cues.

5.2 Overall Discussion

5.2.1 Comparison to previous work

This study has revealed similar problems of ED models as cognitive models as already found by McCurdy (2019): a) the overall distributions of plural classes are perhaps more similar than in McCurdy (2019) but still far from fitting perfectly. Likewise, we

[1] The within-model variability might be only this (comparatively) high because models were tested on two conditions - assuming that each condition elicits a different pattern of productions this would lead to higher variability compared to speakers who were tested on only one condition each. However, variability was only slightly diminished when including each model twice, once for each condition: M 5.5 (SD 0.61).

[2] All correlation coefficients can be found in Appendix C.

see only a slight improvement of correlation between model and speaker data overall. b) the general variability across both speakers and words is considerably lower than in speaker data. Since German plurals cannot be reliably predicted statistically, external factors might be at play that have not been explored so far. The inclusion of gender and (an approximation of) semantics might have enhanced the variability across words and models, but this might also be attributed to the wug words used in the present study.

Results of both speaker and model experiments might be substantially influenced by word-specific effects. Our models did not show the same failure modes as previous work, such as a tendency not to produce *-n* plurals in the models. Furthermore, no statistically reliable effect on the use of *-s* could be attributed to rhyme in either model or speaker data, contrary to Marcus et al. (1995); McCurdy (2019). Since a main difference to the two previous studies was the choice of wug words, our results might differ because they are subject to word-specific effects. Future research with further sets of wug words would therefore be desirable, in order to establish whether the observed patterns in this and previous studies can be replicated or are indeed word-specific.

Corkery et al. (2019) found a tension between higher correlation on less trained models and higher test accuracy on fully trained models for English. This could not be replicated for German, where less trained models showed both higher test accuracy and model correlation. Assuming that fully trained models are more cognitively plausible, this should perhaps make us cautious about drawing final conclusions especially on the issue of variability. On the other hand it calls into question whether models of German plural inflection should indeed be trained to full convergence on the training set. The Wiktionary corpus lists 2.4% of words with at least one alternative plural form; also discussed in McCurdy et al. (2020) in the context of regional variations), questioning whether it is possible/likely for speakers to fully master the German plural system. This should be further investigated for example by examining speech errors in German speakers.

5.2.2 Semantic influences

Starting with results from the literature suggesting that semantics have predictive value for the selection of plural class in German (Williams et al., 2020), we hypothesised that a model of German noun inflection might give more accurate predictions of speakers' use of plurals in wug words when provided with semantic information. Our results suggest that this is not the case. Indeed, the patterns emerging from both experiments point into entirely different directions: while the semantic category *person* had a negative effect on the use of *-e* and a positive one on the *-s* class in the speaker data, in the models there was none on *-e* , a negative one of *-s* and a positive one on *-n* and -∅ . In general, semantics seem to have had a stronger effect on model predictions than on speaker productions (compare the plural class distributions across semantic categories in the two left panels of Figure 5.1). Note, however, that the rather strong effect of semantics in ED models might be only the result of a) our rather coarse way of presenting semantics to the model and b) the inclusion of precisely those categories relevant for the hypothesis, possibly hand-crafting the results into the model. By reducing the number of categories to six, the model is given clues about what might be important information for the task at hand - information the speakers do not have. This could make it more likely for the models to pick up an influence of those categories. A more fine-grained representation of semantics such as image or word embeddings would thus be desirable. Nonetheless, the effect of semantics points in opposite directions in the model and the speaker data. Since semantic category is predictive for the *-n* plural class, as shown in Section 3.1, it is unsurprising that the ED models pick up the relationship. On the other hand, an unexpected negative effect of *person* on *-e* plurals in the corpus data was picked up by speakers, but not by the ED models. This makes it plausible that there additional influences interacting with semantics, which have not been identified and been controlled for in the current study.

Subsequently, even though the model showed a higher correlation to the speaker data than previous studies (McCurdy, 2019), we can only tentatively link this to the inclusion of semantics. At the very least, a higher number of reference models should be trained, in order to test the difference in correlation for statistical reliability. There might be a tendency of the models to show greater variability when provided with more potential cues (e.g. grammatical gender, semantics), but this should be further investigated before drawing final conclusions.

We might conclude that semantics simply do not influence the plural productions of German speakers. This would then set apart German from English, where such an effect has indeed been found (Ramscar, 2002). On the other hand, the German inflection system is generally considerably more complex than English. Together with the high variability across speakers and words, this suggests that German plural productions might be influenced by other factors which have not been identified yet. In future research, these influences should be further investigated, and subsequently included in the model architecture, increasing our understanding of the underlying mechanisms.

Chapter 6

Conclusions

6.1 Findings

Our first question was whether there is an influence of German speaker's plural productions in wug words. To test this, we used the observation by Gaeta (2008), namely that German masculine nouns describing a person are more likely to take an *-n* plural than nouns describing objects. While this observation was verified in corpus data, the effect could not be seen in speaker data. However, there was an unexpected influence of semantics on other classes in the production and the rating data. This suggests a) that participants were aware of the semantics, b) that they do associate endings with semantics, and c) that the effects of semantics might be part of a more complex system of influences from various, as yet unidentified factors on the German plural system.

We can thus conclude that the concrete hypothesis - the influence of the *person* semantic category on the use of *-n* endings - does not hold. However, we cannot finally reject the hypothesis that semantics have an influence on German pluralisation, since effects varied across plural classes and partially contrary to tendencies in corpus data.

Secondly, does the inclusion of semantics help the Encoder-Decoder models to predict production probabilities of plurals in German speakers? We found that influences of semantic category in the model are often opposite to those in the speaker data. The overall correlation between model and speaker data is higher than in previous work (McCurdy, 2019; McCurdy et al., 2020) - but this might be due to a different test set being used - and higher than between reference models trained without semantics and the speaker data.

Our model results indicate a tendency to show more variability the more informative cues are included as inputs. This cannot conclusively be attributed to the semantics, however - other influences such as the words or different training regime might be at work. High variability across words and speakers has been observed in multiple studies (McCurdy, 2019; Zaretsky et al., 2013), yet it seems Encoder-Decoder models are unable to fully account for this. The question remains whether the variability arises from measurable influences on the plural system or whether it is random variation originating in other sources

Our hypothesis necessitated the use of a different set of wug words than the original one used by Marcus et al. (1995); McCurdy (2019). This led to some surprising, additional findings. Most importantly, we did not find an effect of rhyme on the use of *-s* plurals, contradicting the findings of Marcus et al. (1995); McCurdy (2019). This suggests that effects identified in previous work might be due to wug word-specific effects rather than to systematic ones. This implies that the generalisability of effects found on a single set of wug words is questionable, and stresses the importance to replicate past but also our current results with a different set of wug words.

6.2 Final remarks

Over many years, the debate about inflectional systems has largely focused on which model best fits the data. As yet, this has not led to conclusive results, as various models show promising results, but none is able to fully explain the patterns observed in German speakers (Marcus et al., 1995; McCurdy, 2019; McCurdy et al., 2020). Thus, instead of bringing forward more, sometimes contradictory evidence for one model or the other, we suggest to explore what really influences speakers' use of plural classes. While we did not find clear evidence for an influence of semantics in the current study there are many other factors to be explored, such as gender, context, priming effects or even external factors such as education or motivation (see also Zaretsky and Lange (2015)). Whether the influences can be best described by a connectionist, rule- or pattern-based model would then only be a second step, driven by a hopefully more thorough understanding of what mechanisms underlie inflection.

6.3 Limitations and Future work

The present study has mainly focused on a single hypothesis: does the inclusion of rather abstract semantic information (person/object) about a word influence how it is inflected. This gave unclear results. Notably, we did not further investigate phonological or morphological factors (other than rhyme and number of syllables) that might interact with the influence of semantics. These should be further investigated. Furthermore, as can be seen in e.g. Williams et al. (2020), the range of possible influences might be wide, and a different one might be explored. It might be worthwhile to explore the effect of other factors in additional studies in order to better understand influences on plural productions which cannot currently be adequately described by phonology alone.

Moreover, we ended up using a rather coarse representation of semantics: six different categories. Representations such as image or word embeddings might give much more fine-grained predictions befitting more the possibly very fine-grained effects of semantics on speakers' plural productions. While we conducted initial experiments with image embeddings without success, different architectures and/or training regimes as well as increasing the size of the training data could be explored.

Finally, as McCurdy (2019) we only trained our models to produce plurals, a more complete model should include the entire inflectional paradigm. Kirov and Cotterell (2018) used an ED model to jointly learn the entire English verb paradigm and Malouf (2017); Cardillo et al. (2018) found that a simple RNN was able to give good test accuracies on the inflectional paradigms of multiple languages. Using their approach to model wug word inflection could potentially also improve the observed variability across models in German speakers (also in McCurdy (2019)).

Bibliography

Farrell Ackerman, James Blevins, and Robert Malouf. Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. *Analogy in grammar: Form and acquisition*, 54:82, 2009.

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

R Harald Baayen and Fermín Moscoso del Prado Martín. Semantic density and past-tense formation in three Germanic languages. *Language*, pages 666–698, 2005.

R. Harald Baayen, Richard Piepenbrock, and Leon Gulikers. The CELEX lexical database [cd rom]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania, 1995.

Heike Baeskow. *Abgeleitete Personenbezeichnungen im Deutschen und Englischen: kontrastive Wortbildungsanalysen im Rahmen des minimalistischen Programms und unter Berücksichtigung sprachhistorischer Aspekte*, volume 62. Walter de Gruyter, 2011.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

Jean Berko. The child's learning of English morphology. *Word*, 14(2-3):150–177, 1958.

James Blevins, Petar Milin, and Michael Ramscar. The Zipfian paradigm cell filling problem. In *Perspectives on Morphological Organization*, pages 139–158. Brill, 2017.

Franco Alberto Cardillo, Marcello Ferro, Claudia Marzi, and Vito Pirrelli. Deep learning of inflection and the cell-filling problem. *Italian Journal of Computational Linguistics*, 4(1):57–75, 2018.

Georgia-Ann Carter, Faheem Kirefu, Rohit Saxena, and Emelie Van De Vreken. Predicting Noun Gender Using Semantic And Phonological Representations. *Unpublished MSc project report, School of Informatics, University of Edinburgh*, 2019.

Giovanni Cassani, Yu-Ying Chuang, and R Harald Baayen. On the semantics of nonwords and their lexical category. *Journal of Experimental Psychology: Learning, Memory, Cognition*, 46(4):621–637, 2019.

François Chollet et al. Keras. `https://keras.io`, 2015.

Yu-Ying Chuang, Marie-lenka Vollmer, Elnaz Shafaei-Bajestan, Susanne Gahl, Peter Hendrix, and R Harald Baayen. The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behaviour Research Methods*, 2020.

Maria Corkery, Yevgen Matusevych, and Sharon Goldwater. Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3868–3877, 2019.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 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. Association for Computational Linguistics, 2016. doi: 10.18653/v1/W16-2002. URL `https://www.aclweb.org/anthology/W16-2002`.

David Crystal. *I*, pages 234–257. John Wiley Sons, Ltd, 2009. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444302776.ch9`.

Ewa Dabrowska. The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language*, 58(4):931–951, 2008.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

Livio Gaeta. Die deutsche Pluralbildung zwischen deskriptiver Angemessenheit und Sprachtheorie. *Zeitschrift für germanistische Linguistik*, 36(1):74–108, 2008.

Spandana Gella and Frank Keller. An evaluation of image-based verb prediction models against human eye-tracking data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 758–763, 2018.

Spandana Gella, Mirella Lapata, and Frank Keller. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *Proceedings of NAACL-HLT*, pages 182–192, 2016.

Ulrike Hahn and Ramin Charles Nakisa. German inflection: Single route or dual route? *Cognitive Psychology*, 41(4):313–360, 2000.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Maria Heitmeier. Informatics Project Proposal: Modelling the influence of semantics on noun inflection. *Unpublished report, School of Informatics, University of Edinburgh*, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-term Memory. *Neural computation*, 9(8):1735–1780, 1997.

James Hoeffner. Are rules a thing of the past? the acquisition of verbal morphology by an attractor network. In *PROCEEDINGS OF THE FOURTEENTH ANNUAL CONFERENCE OF THE COGNITIVE SCIENCE SOCIETY: JULY 29 TO AUGUST 1,* *1992, COGNITIVE SCIENCE PROGRAM, INDIANA UNIVERSITY, BLOOMINGTON*. Citeseer, 1992.

Marc F. Joanisse and Mark S. Seidenberg. Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences of the United States of America*, 96(13):7592–7, 1999.

Emmanuel Keuleers and Marc Brysbaert. Wuggy: A multilingual pseudoword generator. *Behavior research methods*, 42(3):627–633, 2010.

Christo Kirov and Ryan Cotterell. Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665, 2018.

Klaus-Michael Köpcke. Schemas in german plural formation. *Lingua*, 74(4):303–335, 1988.

Stephan Lewandowsky and Simon Farrell. *Computational modeling in cognition: Principles and practice*. SAGE publications, 2010.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015.

Robert Malouf. Abstractive morphological learning with a recurrent neural network. *Morphology*, 27(4):431–458, 2017.

Gary Marcus, Ursula Brinkmann, Harald Clahsen, Richard Wiese, and Steven Pinker. German inflection: The exception that proves the rule. *Cognitive psychology*, 29(3): 189–256, 1995.

Kate McCurdy. Neural Networks Don't Learn Default Rules for German Plurals, But That's Okay, Neither Do Germans. *Master's Thesis, University of Edinburgh*, 2019.

Kate McCurdy, Sharon Goldwater, and Adam Lopez. Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756. Association for Computational Linguistics, 2020. URL `https://www.aclweb.org/anthology/2020.acl-main.159`.

Ken McRae, George Cree, Mark Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559, 2005.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Ramin Charles Nakisa and Ulrike Hahn. Where defaults don't help: the case of the german plural system. In *Proc. 18th Annu. Conf. Cogn. Sci. Soc*, pages 177–182, 1996.

Robert M Nosofsky. Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 115(1):39, 1986.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Adapting deep network features to capture psychological representations. *arXiv preprint arXiv:1608.02164*, 2016.

Steven Pinker and Alan Prince. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193, 1988.

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

Kim Plunkett and Patrick Juola. A connectionist model of english past tense and plural morphology. *Cognitive Science*, 23(4):463–490, 1999.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Michael Ramscar. The role of meaning in inflection: Why the past tense does not require a rule. *Cognitive Psychology*, 45(1):45–94, 2002.

Daniel Rumelhart and James McClelland. On Learning the Past Tenses of English Verbs. In *Parallel Distributed Processing*, volume 2, page 216–271. MIT Press, 1986.

David Rumelhart, Geoffrey Hinton, and Ronald Williams. Learning by error backpropagation. In *Parallel Distributed Processing*, volume 1. MIT press, 1986.

Dora Schulz and Heinz Griesbach. *Grammatik der deutschen Sprache*. Max Hueber Verlag, München, 11 edition, 1981.

Mark Seidenberg and David Plaut. Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive science*, 38(6):1190–1228, 2014.

Ingrid Sonnenstuhl and Axel Huth. Processing and representation of german-n plurals: A dual mechanism approach. *Brain and Language*, 81(1-3):276–290, 2002.

Ilya Sutskever, Oriol Vinyals, and Quoc Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

Marc Tanti, Albert Gatt, and Kenneth Camilleri. Where to put the image in an image caption generator. *Natural Language Engineering*, 24, 03 2017. doi: 10.1017/S1351324918000098.

Luis Von Ahn and Laura Dabbish. ESP: Labeling Images with a Computer Game. In *AAAI spring symposium: Knowledge collection from volunteer contributors*, volume 2, 2005.

Gert Westermann and Risto Miikkulainen. Verb inflections in German child language: A connectionist account. 1994.

Adina Williams, Tiago Pimentel, Arya McCarthy, Hagen Blix, Eleanor Chodroff, and Ryan Cotterell. Predicting declension class from form and meaning. In *Proceedings* *of the 58th Annual Meeting for the Association of Computational Linguistics*. York, 2020.

Eugen Zaretsky and Benjamin P Lange. No matter how hard we try: Still no default plural marker in nonce nouns in modern high german. In *A blend of MaLT: Selected contributions from the Methods and Linguistic Theories Symposium*, pages 153–178, 2015.

Eugen Zaretsky, Benjamin P Lange, Harald A Euler, and Katrin Neumann. Acquisition of german pluralization rules in monolingual and multilingual children. *Studies in Second Language Learning and Teaching*, 3(4):551–580, 2013.

Appendix A

Stimuli

two syllables		**three syllables**	
rhyme	non-rhyme	rhyme	non-rhyme
Julot	Illtemp	Karagul	Jakaselb
Filast		Pobekus	Rerofept
Silit		Femmotak	Eseruns
Wangom		Zackabat	Klerituld
Junderz		Sketowos	Bolotulz
Toldar		Katulee	
Bagisk		Walgimirz	
Zorimp			
Gnamarz			
Lahrnotz			

Table A.1: Wug words used.

Image	Sentence
img19.jpg	Das ist ein Apfel. Jeder Apfel gehört der Familie der Rosengewächse an. Während dieser recht süß schmeckt, schmecken andere ____ eher bitter.
img4.jpg	Das ist ein Leopard. Dieser Leopard lebt in Marokko. Dieser Leopard frisst vor allem Hirsche, andere ____ jagen aber auch Vögel.
img50.jpg	Links auf dem Bild ist ein Katheter zu sehen. Dieser wird verwendet, um Hohlorgane zu befüllen oder zu entleeren. Dieser Katheter ist aus Gummi, andere ____ sind aus Glas, Silikon oder Kunststoff.
img54.jpg	Das ist ein Vulkan. Dieser Vulkan ist entstanden, weil Magma an die Oberfläche des Erdballs gestiegen ist. Während dieser Vulkan raucht, sind andere ____ immer still.
img81.jpg	Hier sieht ist ein Anzug zu sehen. Dieser Anzug wird zu formellen Anlässen, wie z.B. Theaterbesuchen, getragen. Dieser Anzug ist schwarz, manche ____ sind aber auch blau.
img13.jpg	Hier ist ein Gorilla zu sehen. Jeder Gorilla hat Arme, die länger als seine Beine sind. Dieser Gorilla weist eine Braunfärbung auf, andere ____ zeigen jedoch auch eine Schwarzfärbung.

Table A.2: Sentences and words used in the practice trials.

Image	Sentence
4002.jpg	Hier liegt ein **wug**. Nicht jeder **wug** ist silbern, aber alle sind aus Metall. Von der Firma, von der dieser **wug** produziert wurde, werden auch noch *weitere* ____ produziert.
6481.jpg	Das ist ein **wug**. Dieser **wug** kann mit einer Hand an der linken Schraube gedreht werden. Obwohl dieser **wug** schwarz ist, sind die meisten *anderen* ____ grün.
8249.jpg	Ein **wug** steht im Dunkeln. Wenn es völlig dunkel ist, kann dieser **wug** leuchten. Dieser **wug** ist kombinierbar, *weitere* ____ können darauf gesteckt werden.
8294.jpg	Hier steht ein **wug** zum Verkauf. Dieser **wug** wird verwendet, um Opas zu transportieren. Allein ist dieser **wug** sehr teuer, wenn man *mehrere* ____ kauft, bekommt man eine Vergünstigung.
img15.jpg	Das ist ein **wug**. Obwohl dieser **wug** aussieht, als wäre er aus Schokolade, ist er in Wirklichkeit aus Rasierschaum. Ein solcher **wug** kann sonntags zum Rasieren verwendet werden und wenn man *mehrere* ____ kombiniert, ist das Ergebnis besonders weich.
img37.jpg	Hier ist ein **wug** zu sehen. Ein solcher **wug** wird meistens zur Stabilisierung von Schränken verwendet. Wenn dieser **wug** nicht ausreicht, können *zwei* ____ kombiniert werden, um besonders stabile Schränke zu erhalten.
img59.jpg	Das ist ein **wug**. Dieser ist ein ganz besonderes Exemplar, weil er grün ist. In dem Museum, in dem sich dieser **wug** befindet, sind auch *weitere* ____ zu sehen.
img66.jpg	Das ist ein **wug**. Dieser **wug** ist aus rotem Glas, was eher selten ist. Dieser **wug** ist ein Einzelstück, aber *manche* ____ wurden auch in Massenproduktion hergestellt.'
img75 2.jpg	Hier ist ein **wug** zu sehen. Jeder **wug** ist dreieckig, wobei dieser zusätzlich oben eine Verzierung aufweist. Dieser **wug** gehört einem Mann, in dessen Orchester *mehrere* ____ gespielt werden.
img75.jpg	Das ist ein **wug**. Dieser **wug** kann mithilfe der Antenne und der Knöpfe gesteuert werden. Ein solcher **wug** ist recht selten, aber manchmal sind sogar *mehrere* ____ ausgestellt.
img82.jpg	Hier ist ein **wug** zu sehen. Ein solcher **wug** wächst an Bäumen und kann ausschließlich an Weihnachten gepflückt werden. Ein **wug** ist meistens lila, in anderen Ländern sind [*die*] ____ jedoch auch rot.
img265.jpg	Das ist ein **wug**. Nicht jeder **wug** ist silbern, es gibt auch bronzefarbene. Dieser **wug** gibt Töne von sich, aber *alle* ____ leuchten, wenn sie neben einer Maschine stehen.

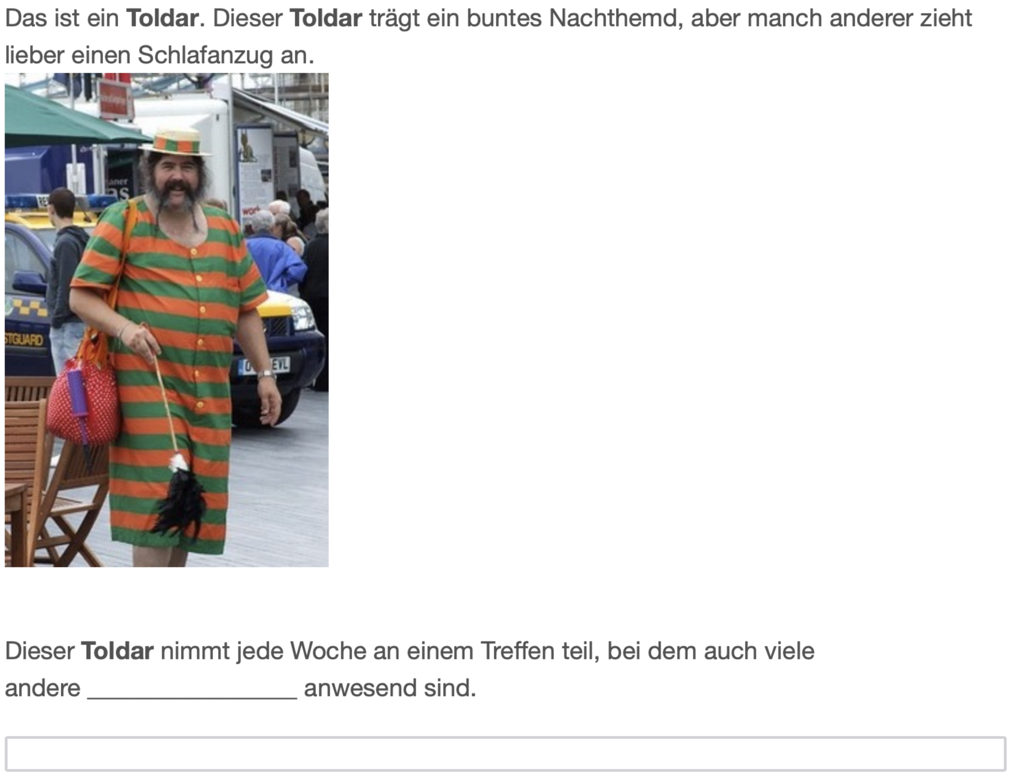
Table A.3: Sentences describing objects. **wug** were replaced with a random wug word (see Table A.1). Italic words were used to elicit plurals in the rating task (see Figure B.2).

Image	Sentence
img1.jpg	Auf diesem Bild ist ein **wug**. Jeder **wug** trägt zu seiner Hochzeit einen Hut aus Korken. Auf einer Doppelhochzeit trägt dieser **wug** und der zweite Bräutigam einen solchen Hut. Es heiraten *zwei* ____
img50.jpg	Vor rotem Hintergrund steht ein **wug**. Wenn ein solcher **wug** fotografiert wird, trägt er schwarze Haare und Make-up. Dieser **wug** trifft *weitere* ____ um ein Gruppenfoto zu machen.
img66 2.jpg	Hier ist ein **wug** in schwarz-weiß abgebildet. Zu Karneval trägt ein männlicher **wug** eine bunte Kappe. Am Faschingsumzug im Februar trifft sich dieser **wug** mit Freunden. An dem Faschingsumzug nehmen dann *viele* ____ teil.
img59 2.jpg	Am Gleis steht ein **wug**. Jeder **wug** trägt sonntags am liebsten einen Hut aus Blumen. Dieser **wug** und *andere* ____ sind gerade auf dem Weg zu Freunden.
img76.jpg	Hier ist ein **wug** neben einem Fragezeichen zu sehen. Dieser **wug** hat rosa Blütenhaare, aber es gibt auch welche mit blauen. Wenn noch ein weiterer **wug** dazu kommt, stehen neben dem Fragezeichen *zwei* ____
img121.jpg	Auf diesem Bild sieht man, wie ein **wug** fest nachdenkt. Für jeden **wug** ist es sehr wichtig, dabei eine Brille zu tragen. Dieser **wug** denkt am liebsten im Stehen, *andere* ____ sitzen lieber.
img141.jpg	Das ist ein **wug**. Dieser **wug** trägt ein buntes Nachthemd, aber manch anderer zieht lieber einen Schlafanzug an. Dieser **wug** nimmt jede Woche an einem Treffen teil, bei dem auch viele *andere* ____ anwesend sind.
img80.jpg	Das ist ein **wug**. Nicht jeder **wug** besitzt eine Uhrenkette, aber alle tragen Strohhüte. Dieser **wug** ist gerade auf dem Weg zu einem Treffen, an dem *mehrere* ____ teilnehmen.
img115.jpg	Hier steht ein **wug**. Jeder **wug** ist ein Fan, der Fahnen am Hut trägt. An diesem Spiel nehmen dieser **wug** und noch *andere* ____ teil.
img154.jpg	Ein **wug** hält eine Handtasche. Nicht jeder **wug** besitzt eine Weste aus Fell, aber es gilt als Statussymbol. Dieser **wug** ist deshalb besonders stolz, wenn ihn *andere* ____ besuchen.
img299.jpg	Ein **wug** schützt sich mit einem Handtuch vor der Sonne. Dieser **wug** ist besonders empfindlich und trägt zusätzlich eine Sonnenbrille. Dieser **wug** geht trotzdem gerne in den Garten, *andere* ____ bleiben am liebsten zuhause.
img374.jpg	Hier schreit ein **wug**. Mancher **wug** lässt sich gerne fotografieren, aber dieser nicht. Dieser **wug** ist besonders fotoscheu, *andere* ____ lassen sich sogar mehrmals täglich fotografieren.

Table A.4: Sentences describing persons. **wug** were replaced with a random wug word (see Table A.1). Italic words were used to elicit plurals in the rating task (see Figure B.2).

Appendix B

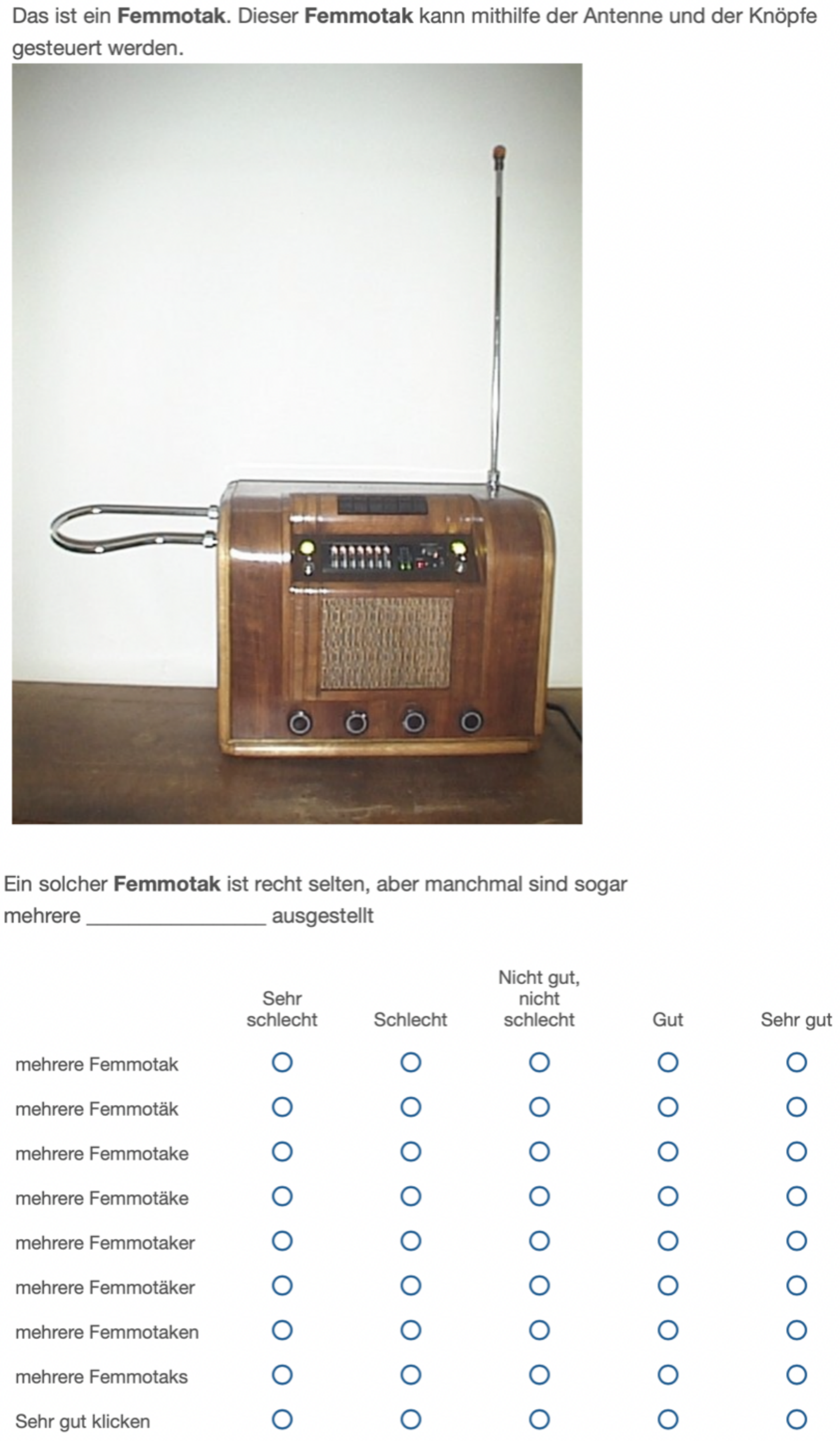
Presentation



Das ist ein **Toldar**. Dieser **Toldar** trägt ein buntes Nachthemd, aber manch anderer zieht lieber einen Schlafanzug an.

Dieser **Toldar** nimmt jede Woche an einem Treffen teil, bei dem auch viele andere ________________ anwesend sind.

Figure B.1: Example of production task. Participants were asked to give the plural form of the bold word.



Das ist ein **Femmotak**. Dieser **Femmotak** kann mithilfe der Antenne und der Knöpfe gesteuert werden.

Ein solcher **Femmotak** ist recht selten, aber manchmal sind sogar mehrere ________________ ausgestellt

	Sehr schlecht	Schlecht	Nicht gut, nicht schlecht	Gut	Sehr gut
mehrere Femmotak	○	○	○	○	○
mehrere Femmotäk	○	○	○	○	○
mehrere Femmotake	○	○	○	○	○
mehrere Femmotäke	○	○	○	○	○
mehrere Femmotaker	○	○	○	○	○
mehrere Femmotäker	○	○	○	○	○
mehrere Femmotaken	○	○	○	○	○
mehrere Femmotaks	○	○	○	○	○
Sehr gut klicken	○	○	○	○	○

Figure B.2: Example of rating task. Participants were asked to rate plural forms of the bold word. When prompted with *Sehr gut klicken* (Click very good) they were asked to select the given rating.

Appendix C

Statistical models

Table C.1: Effect of semantic category on use of plural classes in masculine nouns in corpus data.

	Dependent variable:				
	n	e	zero	s	er
	(1)	(2)	(3)	(4)	(5)
person	1.356^{***}	-1.931^{***}	0.967^{***}	-0.472^{***}	0.126
	(0.072)	(0.069)	(0.056)	(0.112)	(0.271)
Constant	-2.216^{***}	0.103^{***}	-0.973^{***}	-2.421^{***}	-4.697^{***}
	(0.055)	(0.033)	(0.037)	(0.060)	(0.172)
Observations	6,008	6,008	6,008	6,008	6,008
Log Likelihood	−2,576.333	−3,505.897	−3,767.015	−1,525.915	−322.116
Akaike Inf. Crit.	5,156.667	7,015.794	7,538.030	3,055.831	648.232
Note:					*p<0.1; **p<0.05; ***p<0.01

Table C.2: Generalised mixed model of plural class production probabilities by 197 German speakers (note that the *-er* model did not converge)

	Dependent variable:					
	n	e	s	er	zero	other
	(1)	(2)	(3)	(4)	(5)	(6)
age	−0.006	−0.001	0.004	0.008	−0.012	0.010
	(0.012)	(0.009)	(0.010)	(0.054)	(0.040)	(0.008)
degree1	−0.392	0.461***	−0.258	0.445	1.933**	−0.052
	(0.247)	(0.174)	(0.217)	(1.167)	(0.843)	(0.159)
country1	−0.193	0.464*	−0.230	0.483	0.267	−0.104
	(0.385)	(0.275)	(0.333)	(2.142)	(1.294)	(0.246)
rhyme1	−0.592	−0.115	−0.452	−1.476***	0.631	0.362
	(0.484)	(0.674)	(1.520)	(0.492)	(0.801)	(0.495)
person1	0.201*	−0.403***	0.327**	0.895	0.373	0.073
	(0.111)	(0.101)	(0.142)	(0.587)	(0.355)	(0.156)
three_syll1	−0.757*	0.124	−0.663	−0.159	0.429	0.582
	(0.429)	(0.590)	(1.326)	(0.666)	(0.735)	(0.441)
person1:three_syll1	−0.189	0.500***	−0.477**	−1.203	−0.350	−0.093
	(0.164)	(0.140)	(0.216)	(0.836)	(0.512)	(0.209)
Constant	−0.574	−0.990	−3.517**	−10.289***	−11.032***	−3.519***
	(0.722)	(0.810)	(1.686)	(2.808)	(2.178)	(0.625)
Observations	4,751	4,751	4,751	4,751	4,751	4,751
Log Likelihood	−2,045.926	−2,640.084	−1,217.341	−136.848	−336.715	−1,324.078
Akaike Inf. Crit.	4,111.852	5,300.168	2,454.682	293.695	693.430	2,668.156
Bayesian Inf. Crit.	4,176.513	5,364.829	2,519.343	358.356	758.091	2,732.817
Note:					*p<0.1; **p<0.05; ***p<0.01	

Table C.3: Linear mixed model of z-scaled ratings of plural classes by 197 German speakers

	Dependent variable:				
	z_rating				
	n	er	s	e	zero
age	−0.001	0.006^{**}	0.001	0.001	-0.006^{*}
	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)
degree1	−0.002	−0.074	-0.216^{**}	0.120^{**}	0.059
	(0.068)	(0.053)	(0.092)	(0.051)	(0.066)
country1	-0.259^{**}	0.017	−0.177	0.126	0.099
	(0.107)	(0.082)	(0.144)	(0.080)	(0.103)
rhyme1	0.048	-0.177^{**}	0.096	−0.178	−0.125
	(0.119)	(0.080)	(0.152)	(0.232)	(0.144)
person1	0.001	0.276^{***}	0.002	0.042	0.013
	(0.055)	(0.043)	(0.068)	(0.055)	(0.053)
three_syll	−0.081	-0.197^{***}	0.093	−0.172	−0.056
	(0.110)	(0.074)	(0.144)	(0.207)	(0.130)
person1:three_syll	−0.051	-0.120^{**}	−0.006	−0.090	−0.083
	(0.078)	(0.059)	(0.101)	(0.073)	(0.072)
Constant	0.800^{***}	-0.511^{***}	0.206	0.409	0.106
	(0.189)	(0.138)	(0.245)	(0.273)	(0.206)
Observations	1,557	2,520	978	2,329	2,520
Log Likelihood	−1,901.900	−2,896.023	−1,241.449	−3,056.373	−3,407.056
Akaike Inf. Crit.	3,825.799	5,814.045	2,504.897	6,134.745	6,836.112
Bayesian Inf. Crit.	3,884.655	5,878.197	2,558.638	6,198.030	6,900.264
Note:					*p<0.1; **p<0.05; ***p<0.01

Table C.4: Linear mixed model of z-scaled ratings of plural classes by 197 German speakers, excluding the two faulty items *Walgimirz* and *Zackabat*

	Dependent variable:				
	z_rating				
	n	er	s	e	zero
age	−0.001	0.006**	0.002	0.002	−0.006*
	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)
degree1	−0.011	−0.071	−0.205**	0.114**	0.052
	(0.068)	(0.053)	(0.092)	(0.052)	(0.066)
country1	−0.253**	0.052	−0.158	0.140*	0.092
	(0.107)	(0.083)	(0.145)	(0.081)	(0.104)
rhyme1	0.065	−0.233***	0.066	−0.289	−0.151
	(0.128)	(0.076)	(0.163)	(0.236)	(0.154)
person1	−0.002	0.281***	0.003	0.045	0.013
	(0.055)	(0.043)	(0.069)	(0.055)	(0.053)
three_syll	−0.063	−0.267***	0.084	−0.324	−0.078
	(0.122)	(0.074)	(0.158)	(0.217)	(0.142)
person1:three_syll	−0.057	−0.107*	−0.054	−0.035	−0.092
	(0.083)	(0.061)	(0.106)	(0.076)	(0.076)
Constant	0.787***	−0.495***	0.199	0.491*	0.140
	(0.194)	(0.135)	(0.252)	(0.273)	(0.212)
Observations	1,424	2,325	916	2,134	2,325
Log Likelihood	−1,748.476	−2,666.546	−1,169.404	−2,799.118	−3,164.265
Akaike Inf. Crit.	3,518.952	5,355.092	2,360.808	5,620.237	6,350.529
Bayesian Inf. Crit.	3,576.826	5,418.359	2,413.828	5,682.560	6,413.795
Note:					*p<0.1; **p<0.05; ***p<0.01

Table C.5: Generalised linear mixed model of 100 aggregated ED model production probabilities

	Dependent variable:					
	n	e	s	zero	er	other
	(1)	(2)	(3)	(4)	(5)	(6)
person1	0.623***	−0.034	−1.412***	2.840***	1.186	3.726***
	(0.116)	(0.106)	(0.148)	(0.740)	(1.194)	(1.026)
three_syll1	−2.323**	0.193	−2.262	3.159*	−0.180	5.334***
	(1.182)	(1.211)	(2.780)	(1.749)	(4.236)	(1.360)
rhyme1	−1.539	−1.539	2.439	0.748	−1.595	0.388
	(1.350)	(1.388)	(2.850)	(1.846)	(4.122)	(1.031)
person1:three_syll1	−0.234	−0.262	0.067	−1.281*	−0.596	−4.513***
	(0.220)	(0.171)	(0.277)	(0.763)	(1.425)	(1.046)
Constant	−1.381	1.342	−7.257**	−9.711***	−12.432**	−9.377***
	(1.449)	(1.497)	(3.332)	(2.163)	(4.959)	(1.516)
Observations	4,800	4,800	4,800	4,800	4,800	4,800
Log Likelihood	−1,388.686	−1,903.618	−941.812	−590.513	−58.591	−560.919
Akaike Inf. Crit.	2,791.371	3,821.235	1,897.624	1,195.027	131.181	1,135.838
Bayesian Inf. Crit.	2,836.706	3,866.570	1,942.959	1,240.361	176.516	1,181.173
Note:						*p<0.1; **p<0.05; ***p<0.01

Table C.6: Generalised linear mixed model of 10 fully trained ED model production probabilities (note that the *-er* , *other* and -∅ did not converge.

	Dependent variable:				
	n	e	s	er	zero
	(1)	(2)	(3)	(4)	(5)
person	1.086***	−0.488	−1.222***	14.616***	1.683***
	(0.397)	(0.341)	(0.454)	(0.004)	(0.001)
three_syll1	−2.656**	0.375	−3.615	13.049***	2.050***
	(1.113)	(1.298)	(2.609)	(0.004)	(0.001)
rhyme1	−1.982*	−1.726	2.068	−1.885***	0.922***
	(1.052)	(1.467)	(2.897)	(0.004)	(0.001)
person:three_syll1	−0.249	−0.171	1.529*	−14.571***	−0.903***
	(0.773)	(0.537)	(0.911)	(0.004)	(0.001)
Constant	−0.910	1.642	−5.812*	−30.862***	−7.400***
	(1.134)	(1.583)	(3.522)	(0.004)	(0.001)
Observations	480	480	480	480	480
Log Likelihood	−143.298	−204.639	−116.987	−12.524	−88.778
Akaike Inf. Crit.	300.596	423.278	247.973	39.049	191.557
Bayesian Inf. Crit.	329.812	452.495	277.190	68.265	220.773
Note:					*p<0.1; **p<0.05; ***p<0.01

Table C.7: Generalised linear mixed model of 100 aggregated ED model production probabilities, replacing semantic and syllable predictors with whether the last noun in the description sentence had the same plural class as the one chosen. Note that *other* did not occur in the sentences and is thus not included here, and the model for *-er* did not converge.

	Dependent variable:				
	n	e	s	er	zero
	(1)	(2)	(3)	(4)	(5)
n_last1	−0.027				
	(0.096)				
e_last1		−0.037			
		(0.078)			
s_last1			0.531^{*}		
			(0.319)		
er_last1				0.350	
				(0.884)	
age	−0.006	−0.001	0.004	0.007	
	(0.012)	(0.009)	(0.011)	(0.143)	
zero_last1					0.133
					(0.307)
degree1	−0.391	0.458^{***}	−0.260	0.444	1.862^{**}
	(0.247)	(0.173)	(0.221)	(2.871)	(0.819)
country1	−0.193	0.462^{*}	−0.237	0.500	0.243
	(0.385)	(0.274)	(0.339)	(5.829)	(1.295)
rhyme1	−0.209	−0.281	−0.080	-1.058^{**}	0.501
	(0.482)	(0.625)	(1.504)	(0.445)	(0.730)
Constant	-1.182^{*}	−0.863	-4.153^{***}	-10.465^{*}	-11.004^{***}
	(0.649)	(0.646)	(1.390)	(6.187)	(1.758)
Observations	4,751	4,751	4,751	4,751	4,751
Log Likelihood	−2,049.361	−2,648.566	−1,219.094	−139.664	−337.290
Akaike Inf. Crit.	4,114.721	5,313.132	2,454.187	295.328	688.579
Bayesian Inf. Crit.	4,166.450	5,364.861	2,505.916	347.056	733.842
Note:				*p<0.1; **p<0.05; ***p<0.01	

Table C.8: Correlation between productions of 10 fully trained ED models and the speaker data. ‘nan’ indicates that the plural class was missing in either the model or speaker data. $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

pl_class	overall
n	0.5*
e	0.55**
zero	0.4
uml_e	0.47
other	0.4
s	0.85***
er	-0.24
uml_er	-1.0***
uml_zero	nan