



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

## Conditioning, but on which distribution? Grammatical gender in German plural inflection

### Citation for published version:

McCurdy, K, Lopez, A & Goldwater, S 2020, Conditioning, but on which distribution? Grammatical gender in German plural inflection. in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics (ACL), pp. 59-65, Workshop on Cognitive Modeling and Computational Linguistics 2020, Virtual Workshop, 19/11/20. <https://doi.org/10.18653/v1/2020.cmcl-1.8>

### Digital Object Identifier (DOI):

[10.18653/v1/2020.cmcl-1.8](https://doi.org/10.18653/v1/2020.cmcl-1.8)

### Link:

[Link to publication record in Edinburgh Research Explorer](#)

### Document Version:

Peer reviewed version

### Published In:

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics

### General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



# Conditioning, but on which distribution? Grammatical gender in German plural inflection

Kate McCurdy Adam Lopez Sharon Goldwater

Institute for Language, Cognition and Computation

School of Informatics

University of Edinburgh

kate.mccurdy@ed.ac.uk

## Abstract

Grammatical gender is a consistent and informative cue to the plural class of German nouns. We find that neural encoder-decoder models learn to rely on this cue to predict plural class, but adult speakers are relatively insensitive to it. This suggests that the neural models are not an effective cognitive model of German plural formation.

## 1 Introduction

In recent years, neural models of natural language have proven to be powerful *statistical learners*, capable of representing linguistic patterns and the conditions under which they generalize to new forms (e.g. Kirov and Cotterell, 2018). Artificial language learning experiments show that humans are also statistical learners: when patterns appear consistently with certain cues in the input, speakers consistently rely on those cues to generalize patterns to new forms (Newport, 2016).

Our research examines how two different statistical learners — neural encoder-decoder (ED) models and adult German speakers — use the cue of grammatical gender in plural inflection of novel words. Gender has a high statistical association with plural suffix: the feminine noun *Wahl* (“vote”) is *Wahlen* in the plural, but the rhyming neuter noun *Mal* (“time”) has the plural form *Male*. We expect that both speakers and the ED model will produce distributions over plural forms which are heavily conditioned on the gender of the input word. We find that the neural model is highly sensitive to grammatical gender; however, speaker productions, while slightly influenced by gender, appear more consistent with a distribution over plural suffixes which is *unconditioned* on gender. This surprising result suggests that, even though gender is a very informative cue to plural class, speakers may preferentially attend to different cues.

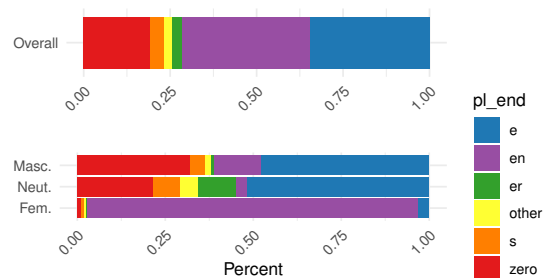


Figure 1: Distribution of plural suffix overall (upper) and by gender (lower) in the UniMorph corpus.

## 2 Background

German plural inflection is realized by five major suffixes, none of which commands a majority in either type or token frequency.<sup>1</sup> The two most frequent suffixes, -e and -(e)n, each apply to roughly 35-40% of German nouns (Figure 1, upper). With no majority class, how can a learner determine which pattern to generalize to new words?

One major cue to plural class comes from a noun’s *grammatical gender*. The strong statistical association between grammatical gender and plural inflection class is widely recognized in the literature (Zaretsky et al., 2013; Yang, 2016; Williams et al., 2020, to cite some recent examples), and readily apparent in Figure 1 (lower), which shows the distribution of plural suffixes by noun gender in the UniMorph corpus (Kirov et al., 2016). Of the two most frequent plural suffixes, -(e)n is highly associated with feminine nouns, and -e with non-feminine (masculine and neuter); the tendency is so strong that some researchers have analyzed these suffixes as gender-conditioned “defaults” (Indefrey, 1999; Laaha et al., 2006). From this perspective, grammatical gender provides a highly consistent cue to plural class membership, which ought to

<sup>1</sup>For simplicity, this discussion focuses on suffixes and omits related phenomena such as umlaut.

inform a statistical learner’s generalizations. Furthermore, grammatical gender is expressed on the article preceding a noun, and this initial position is perceptually salient to speakers (Frigo and McDonald, 1998). These properties suggest that grammatical gender should influence how speakers inflect novel nouns, and indeed, several *wug tests*<sup>2</sup> on German-speaking adults have found significant effects of gender (Köpcke 1988; Zaretsky and Lange 2016; though c.f. Marcus et al. 1995). Based on the artificial language learning literature (e.g. Newport, 2016), we might expect speakers to display *conditional probability matching* on novel German nouns, such that the probability of a noun taking a certain plural inflection — in particular, the two highly frequent classes -e and -(e)n — depends upon its grammatical gender.

Neural encoder-decoder (ED) models have recently been proposed for consideration as models of speaker cognition (Kirov and Cotterell, 2018). This has prompted investigation into the extent to which these models capture speaker behavior (Corkery et al., 2019; King et al., 2020; McCurdy et al., 2020). Earlier work suggests that neural models of German plural inflection are sensitive to grammatical gender: Goebel and Indefrey (2000) found that a simple recurrent network learned to favor -e plurals for masculine nouns, and -(e)n when the same nouns were presented as feminine gender. We hypothesize that neural models and adult speakers are equally capable of using the information available from grammatical gender to predict number inflection. We expect both to demonstrate similar gender-conditioned probability matching to the distribution shown in Figure 1 (lower), resulting in a majority use of -(e)n for feminine nouns, and -e for masculine and neuter nouns.

### 3 Method

To compare how grammatical gender influences plural inflection for German speakers and neural models, we use a parallel production task on nonce words (a *wug test*) for both speakers and model. Our study largely follows the data collection and modeling procedures of McCurdy et al. (2020).

**Stimuli** We use the 24 made-up nouns developed by Marcus et al. (1995), listed in Appendix A. By design, these nouns lack strong phonological cues

<sup>2</sup>A *wug test* is the task of inflecting a novel word (e.g. “wug” in English). This task lets researchers observe which inflectional variants speakers use (e.g. plural “wugs”; Berko, 1958).

to plural class.<sup>3</sup> In their original study, Marcus et al. did not find a significant effect of grammatical gender; however, Zaretsky and Lange (2016) used the same stimuli and reported gender effects in the expected direction — participants used -(e)n more on feminine nouns, and -e more for nonfeminine nouns. Zaretsky and Lange speculate that these discrepant findings stem from differences in the two study designs: scale (the earlier study had 48 participants, the later one 585) and task (acceptability ratings vs. elicited productions). A third differentiating factor is the presence of semantic cues in the Marcus et al. study, which provided sentence contexts around the nonce words; for example, a sentence like *Die grünen BRALS sind billiger* (“The green brals are cheaper”) would imply that the nonce word *Bral* referred to an object, whereas *Die BRALS sind ein bißchen komisch* (“The Brals are a bit weird”) would imply that *Bral* was a family name. As adult learners can attend to formal and semantic cues under different conditions (Culbertson et al., 2017), it’s possible that this manipulation directed participant focus toward semantic cues rather than grammatical gender. Zaretsky and Lange provided no semantic context in their experiment, only presenting the indefinite article and word form to participants (e.g. *Ein Bral*, “a [masculine/neuter] bral”). Our experimental design for both speakers and the neural model is closer to that of Zaretsky and Lange (2016): we elicit plural form productions and provide no semantic cues. This suggests we might also expect to find a robust effect of grammatical gender for these stimuli.

**Human data collection** We collected production data from 92 native German speakers<sup>4</sup> through an online survey. Participants were shown the singular form of each noun preceded by a definite article indicating grammatical gender (e.g. *Der Bral* for masculine, *Das Bral* neuter, *Die Bral* feminine), and prompted to type a plural-inflected form for each of the 24 nouns. Participants were randomly assigned to one of three lists. Grammatical gender was counterbalanced within each list, such that each participant saw 8 feminine, 8 masculine, and 8 neuter nouns. Gender was also counterbalanced

<sup>3</sup>Certain phones in word-final position are highly associated with specific plural suffixes. The most consistent associations are found with vowels: words ending in schwa generally take -(e)n, and words ending in full vowels generally take -s (MacWhinney, 1978). Our stimuli all have word-final consonants, which lack strong associations with plural class.

<sup>4</sup>Participants were recruited through the platform Prolific. Of 100 tested, 8 were excluded for failing attention checks.

across lists, such that each noun appeared with a different gender in each list.

**Encoder-decoder model** A neural encoder-decoder (ED) model encodes an input sequence into a fixed vector representation and then incrementally decodes it into a corresponding output sequence (Sutskever et al., 2014). We follow other recent work in using the architecture of Kann and Schütze (2016), which has been proposed for cognitive modeling (Kirov and Cotterell, 2018).

For the task of German number inflection, the ED takes as input a character sequence representing the singular nominative form of a noun, preceded by a special character for grammatical gender (e.g.  $\langle f \rangle$  W A H L;  $\langle f \rangle$  indicates feminine,  $\langle m \rangle$  masculine, and  $\langle n \rangle$  neuter). The model is trained to produce the noun’s corresponding nominative plural form as output (e.g. W A H L E N). We used the 11,243 German nouns in UniMorph (Kirov et al., 2016) as our corpus, and added noun gender by merging the dataset with another Wiktionary scrape.<sup>5</sup> We follow the modeling procedure of McCurdy et al. (2020), who found that neural ED models correctly learned the most frequent plural suffix for neuter stimuli, but did not evaluate sensitivity to grammatical gender; please see their paper for further implementation details.

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

## 4 Results

Our results (Figure 2) show that both speakers and the ED model are sensitive to grammatical gender, but the model relies on this cue considerably more than speakers. Statistical analysis confirms that a) both speakers and the model show reliable effects of grammatical gender on their plural form productions, and b) gender effects are substantially greater for model productions. We fit two logistic mixed-effects models to separately analyze production of -e and -(e)n. Details of our analysis can be found in Appendix B. For both -e and -(e)n production,

<sup>5</sup><https://github.com/gambolputty/german-nouns/>

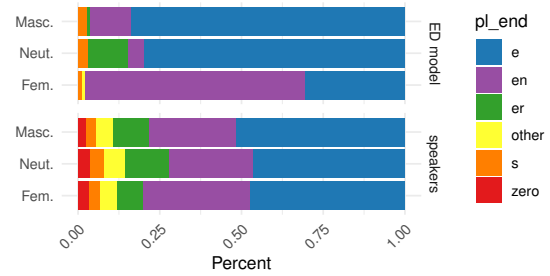


Figure 2: Plural suffix productions by gender, speakers (upper) vs. ED model (lower)

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

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

we found a significant main effect of gender, and a significant interaction with data source (model or speakers), indicating that gender effects were amplified by the model.

Intriguingly, the speaker productions are not only less sensitive to grammatical gender, they also appear very consistent with the overall type frequency distribution of the plural suffixes, *unconditioned* on gender. To quantify this intuition, we looked at how the distribution of plural suffixes produced over each of the 72 noun-gender item combinations *correlated* to various other metrics. We asked three questions: 1) How well do item-level speaker and ED model productions correlate with *each other*? 2) How well do both sets of item-level productions correlate with the *gender-conditioned distribution* of plural suffix types observed in the German lexicon? 3) How well do both sets correlate with the *unconditioned overall distribution* of types? Table 1 shows the results: while item-level ED outputs are most correlated with the gender-conditioned distribution, item-level speaker data is most correlated<sup>6</sup> with the overall (unconditioned) type frequency.<sup>7</sup> Even though the speaker and ED data are

<sup>6</sup>Table 1 shows results from Pearson’s linear correlation; analysis with Spearman’s rank correlation coefficient showed the same trend.

<sup>7</sup>To avoid potential confounds from both training and measuring on the UniMorph corpus, our estimates of gender-conditioned and overall type frequency are derived from Zaret-

matched by item, their productions have a lower correlation with each other than with the general type-frequency distributions.

## 5 Discussion

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

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

One possibility is that the phonological forms of our noun stimuli provide their own statistical conditioning, to a stronger degree than anticipated. This is illustrated in Figure 3, which plots the distribution of nouns in UniMorph sharing two key properties with our stimuli: they are monosyllabic and end in a consonant. On the one hand, nouns with this type of form clearly also show gender conditioning, with -(e)n much more prevalent among feminine

sky et al. (2013)’s analysis of the thousand most frequent nouns from the DeReWo corpus.

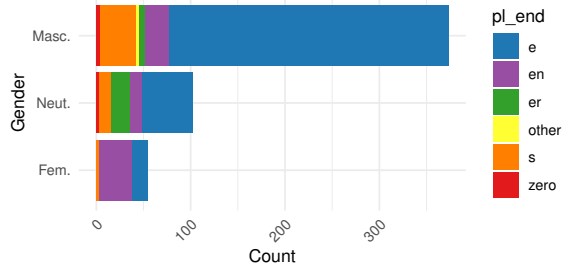


Figure 3: Distribution of plural suffix by gender for consonant-final monosyllabic words in UniMorph

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

Table 2: Correlations (Pearson’s  $r$ , 95% CI in parens) between item-level production percentages for speakers and ED model with 1) overall type frequency (Overall-TF), 2) gender-conditioned type frequency (Gender-TF), *only* considering consonant-final monosyllabic nouns in UniMorph (shown in Figure 3).

nouns. On the other hand, nouns with this general form are predominantly masculine gender, and the numerical prevalence of nonfeminine forms may diminish speakers’ sensitivity to a rare feminine gender cue, such as they encounter in our experiment. Under this account, adult speakers condition their plural productions upon phonological form to a greater extent than grammatical gender. The results in Table 2 further support this interpretation. Looking only at the consonant-final monosyllabic words plotted in Figure 3, ED model productions show a higher correlation to the gender-conditioned distribution over plural suffixes, while the highest correlation generally (.78) appears between speakers productions and the overall distribution of plural classes for these phonologically similar words. The potential shortcoming of the ED as a cognitive model, then, is that it assigns too much weight to the cue of grammatical gender, even though it is statistically reasonable to do so.

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

## References

- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. [Fitting Linear Mixed-Effects Models Using lme4](#). *Journal of Statistical Software*, 67(1):1–48.
- Jean Berko. 1958. [The Child’s Learning of English Morphology](#). *WORD*, 14(2-3):150–177.
- Maria Corkery, Yevgen Matusevych, and Sharon Goldwater. 2019. [Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3868–3877, Florence, Italy. Association for Computational Linguistics.
- Jennifer Culbertson, Annie Gagliardi, and Kenny Smith. 2017. [Competition between phonological and semantic cues in noun class learning](#). *Journal of Memory and Language*, 92:343–358.
- Lenore Frigo and Janet L. McDonald. 1998. [Properties of Phonological Markers That Affect the Acquisition of Gender-Like Subclasses](#). *Journal of Memory and Language*, 39(2):218–245.
- Rainer Goebel and Peter Indefrey. 2000. A recurrent network with short-term memory capacity learning the German-s plural. *Models of language acquisition: Inductive and deductive approaches*, pages 177–200.
- Peter Indefrey. 1999. Some problems with the lexical status of nondefault inflection. *Behavioral and Brain Sciences*, 22(6):1025–1025.
- Katharina Kann and Hinrich Schütze. 2016. [Single-Model Encoder-Decoder with Explicit Morphological Representation for Reinflection](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 555–560, Berlin, Germany. Association for Computational Linguistics.
- David L King, Andrea D Sims, and Micha Elsner. 2020. Interpreting Sequence-to-Sequence Models for Russian Inflectional Morphology. In *Proceedings of the Society for Computation in Linguistics*, volume 3.
- Christo Kirov and Ryan Cotterell. 2018. [Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker and Prince \(1988\) and the Past Tense Debate](#). *Transactions of the Association for Computational Linguistics*, 6:651–665.
- Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. [Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3121–3126, Portorož, Slovenia. European Language Resources Association (ELRA).
- Klaus-Michael Köpcke. 1988. [Schemas in German plural formation](#). *Lingua*, 74(4):303–335.
- Sabine Laaha, Dorit Ravid, Katharina Korecky-Kröll, Gregor Laaha, and Wolfgang U. Dressler. 2006. [Early noun plurals in German: regularity, productivity or default?](#) *Journal of Child Language*, 33(2):271–302.
- Brian MacWhinney. 1978. [The Acquisition of Morphophonology](#). *Monographs of the Society for Research in Child Development*, 43(1/2):1.
- Gary F Marcus, Ursula Brinkmann, Harald Clahsen, Richard Wiese, and Steven Pinker. 1995. German inflection: The exception that proves the rule. *Cognitive psychology*, 29(3):189–256.
- Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. [Inflecting When There’s No Majority: Limitations of Encoder-Decoder Neural Networks as Cognitive Models for German Plurals](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756, Online. Association for Computational Linguistics.
- Elissa L. Newport. 2016. [Statistical language learning: computational, maturational, and linguistic constraints](#). *Language and Cognition*, 8(3):447–461.
- R Core Team. 2019. [R: A Language and Environment for Statistical Computing](#). R Foundation for Statistical Computing, Vienna, Austria.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.
- Adina Williams, Tiago Pimentel, Hagen Blix, Arya D. McCarthy, Eleanor Chodroff, and Ryan Cotterell. 2020. [Predicting Declension Class from Form and Meaning](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6682–6695, Online. Association for Computational Linguistics.
- Charles D. Yang. 2016. *The price of linguistic productivity : how children learn to break the rules of language*. The MIT Press, Cambridge, Massachusetts.
- Eugen Zaretsky and Benjamin P Lange. 2016. No matter how hard we try: Still no default plural marker in nonce nouns in Modern High German. In *A blend of MaLT: selected contributions from the Methods and Linguistic Theories Symposium 2015*, number Band 15 in Bamberger Beiträge zur Linguistik, pages 153–178. University of Bamberg Press, Bamberg.
- Eugen Zaretsky, Benjamin P. Lange, Harald A. Euler, and Katrin Neumann. 2013. [Acquisition of German pluralization rules in monolingual and multilingual children](#). *Studies in Second Language Learning and Teaching*, 3(4):551.

Suffix	Fixed effect	Est. $\beta$	Std. Err.	z value	$Pr(>  z )$
-(e)n	(Intercept)	-1.51086	0.24064	-6.278	3.42e-10 ***
	gender.masc	-1.01487	0.08248	-12.304	< 2e-16 ***
	gender.neut	-0.39048	0.06956	-5.613	1.99e-08 ***
	source.ED	-0.09325	0.15534	-0.600	0.548
	gender.m:src.ED	-0.81875	0.08209	-9.974	< 2e-16 ***
	gender.n:src.ED	-0.27620	0.06944	-3.978	6.96e-05 ***
-e	(Intercept)	0.37876	0.18847	2.010	0.044473 *
	gender.masc	0.31680	0.05581	5.676	1.38e-08 ***
	gender.neut	0.58894	0.05727	10.283	< 2e-16 ***
	source.ED	0.43671	0.13094	3.335	0.000852 ***
	gender.m:src.ED	0.42172	0.05586	7.549	4.38e-14 ***
	gender.n:src.ED	0.41830	0.05713	7.322	2.44e-13 ***

Table 3: Summary of fixed effects from logistic regression analysis

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

Table 4: Experimental stimuli (Marcus et al., 1995)

## A Stimuli

Table 4 shows the 24 noun stimuli used in our experiment. In their original study, Marcus et al. distinguished between Rhymes, which rhyme with existing German words, and Non-Rhymes, which don't. As this distinction is not relevant to our study, we omit it from our analysis.

## B Statistical analysis

Here we report the results of our statistical model of the production of -e and -(e)n. We fit two separate mixed-effect binomial logistic models using the lme4 package (Bates et al., 2015) in R (R Core Team, 2019). Item (i.e. stimulus word) and subject (participant for human study, random seed for ED model) were included as random effects. Both models were fit using a stepwise procedure. We started with a baseline model of intercept plus random effects and incrementally added the following fixed

effects (with sum-coded contrasts): grammatical gender (masculine coded as 1, neuter as 2, feminine not contrasted), data source (ED model coded as 1, speakers not contrasted), gender by source interaction. Each additional fixed effect produced a significantly improved fit as measured by a chi-squared test.

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

**Sanity checks** As human speakers show high inter-participant variability on this task (Fig. 4), we performed additional separate analysis on the speaker data.<sup>8</sup> We fit the same model as previously described, with the exception that the data source factor was omitted, as all data came from speakers. We also fit models using Masculine and Neuter as the reference gender in the sum contrast coding scheme, to see whether they yielded different re-

<sup>8</sup>We thank an anonymous reviewer for highlighting this issue and suggesting these validity checks.

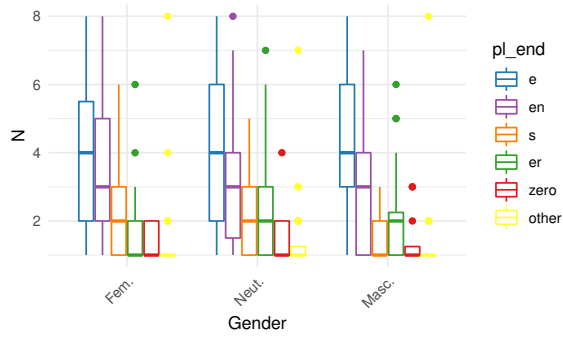


Figure 4: Individual speaker variation in plural suffix production by gender. Each speaker saw 8 words from each gender, shown on the y-axis. For each gender and plural suffix, the boxes indicate the median and interquartile range of individual speaker productions for that combination. For all gender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

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

Table 5: Speaker data only: statistical effect of different gender reference levels in contrast coding. Header shows reference level, cells show estimated coefficient (standard error in parentheses). Estimates for Masculine reference level are identical to rows already shown (e.g. suffix -(e)n:  $-.19$  for gdr.neut,  $.31$  for gdr.fem). Stars indicate significance level: \*  $\leq 0.05$ , \*\*  $\leq 0.01$ , \*\*\*  $\leq 0.001$ .

sults from the original model's Feminine reference level.

The speaker-only model shows a reduced but consistent effect of gender (Tab. 5). Speakers reliably produce -(e)n more for feminine nouns, and less for neuter nouns, relative to the grand mean. Speakers also reliably produce -e more for masculine nouns. These difference are statistically significant even though, for all three genders, speakers produce -e more than -(e)n (Fig. 4).