Modeling grammatical gender and plural inflection in German

Kate McCurdy, Adam Lopez, Sharon Goldwater (University of Edinburgh)
kate.mccurdy@ed.ac.uk

If a German speaker needs to produce the plural form of an unknown word such as *Bral*, that speaker must decide whether *Bral* belongs to the same inflectional class as *das Mal* (yielding an /-e/ suffix for plural *Brale*), *die Wahl* (yielding /-(e)n/: *Bralen*), or another class (see linguistic addendum). Grammatical gender provides a potential clue: for example, if the novel word is feminine (*die Bral*), it might follow *die Wahl* and take the /-(e)n/ suffix like most feminine nouns (c.f. Fig. 1). Researchers have found that that grammatical gender shares substantial mutual information with plural class (Williams et al., 2020), and influences how adult speakers inflect novel nouns (Köpcke, 1988; Zaretsky and Lange, 2016; though see Marcus et al., 1995).

Neural encoder-decoder (ED) networks 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; McCurdy et al., 2020). If neural models of German plural inflection are indeed sensitive to grammatical gender in a similar way to speakers (as earlier findings suggest, c.f. Goebel and Indefrey, 2000), this might support their application as cognitive models. The current study aims to evaluate whether ED models show a speaker-like response to grammatical gender, by presenting speakers and models with the same production task on the same novel stimuli. As existing data resources were insufficient for fine-grained comparison, we collected speaker production data to support item-level analysis. Based on the empirical distribution of suffixes by gender, we expected both speakers and the model would prefer /-(e)n/ for feminine nouns and /-e/ for nonfeminine.

Following Zaretsky and Lange (2016), we use the novel nouns developed by Marcus et al. (1995) as our evaluation stimuli (Tab. 1). Participants in our online survey were shown the singular form of each noun preceded by a definite article indicating grammatical gender (e.g. *Die Bral*), and prompted to type a plural-inflected form for each of the 24 nouns. Participants were randomly assigned to one of three lists counterbalanced for gender: each list contained 8 masculine, 8 neuter, and 8 feminine nouns, and each noun appeared with a different gender across lists. We tested 100 participants with this setup in Phase 1 of the experiment, excluding 8 for failed attention checks. The first 92 participants appeared surprisingly insensitive to grammatical gender, so in Phase 2, we added bonus rewards; Perfors (2016) found that a similar incentive scheme increased adult speakers' tendency to regularize inconsistent morphology, so we expected this would motivate closer attention to salient cues such as gender. We offered participants a 2 cent bonus for each plural form production which matched the form produced by a majority of other speakers (calculated from Phase 1 data). 100 participants were tested in Phase 2. As their behavior showed no statistical difference from Phase 1 participants, we combine all participant data in our analysis. To compare with neural models, we train Kann and Schütze's Morphological Encoder-Decoder (2016) on the UniMorph corpus (Kirov et al., 2016; Fig. 1). We trained one model with grammatical gender cues and a baseline model on word form alone with no gender. Please see the technical addendum for details.

Fig. 2 shows the results: the neural model is far more sensitive to grammatical gender than speakers on this task. Statistical analysis of /-e/ and /-(e)n/ production (Tab. 2) found a significant main effect of gender, indicating that it influences both speaker and model productions; however, both analyses also found significant interactions with data source (ED model with gender vs. speakers), indicating that model productions were more sensitive to gender. In fact, item-level speaker productions are more correlated to those of the baseline ED model *without* gender than the model with gender (Tab. 3). This surprising result suggests speakers may attend more to word form than to gender. By contrast, neural models learn to rely on grammatical gender, and behave more like speakers when this cue is removed from the input.

Additional material

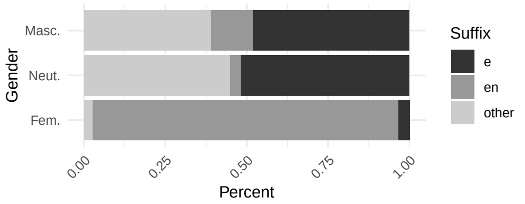

Figure 1: Plural class distribution by gender in the UniMorph corpus (Kirov et al., 2016), which was used to train the ED models. The dominant suffix for feminine nouns is /-(e)n/, while /-e/ is used for the plurality of non-feminine nouns. See Tab. 4 for other suffixes.

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

Table 1: Experimental stimuli (Marcus et al., 1995). The original experiment's Rhyme / Non-Rhyme distinction is not relevant for us.

References

Bahdanau et al., ICLR, 2015
Bates et al., J. of Stat. Software, 2015
Corkery et al., ACL, 2019
Goebel & Indefrey, Models of Lang. Acq., 2000
Kann & Schütze, ACL, 2016
Kirov & Cotterell, TACL, 2018
Kirov et al., LREC, 2016
Köpcke, Lingua, 1988
Marcus et al., Cog. Psych., 1995
McCurdy et al., ACL, 2020
Perfors, Lang. Learning and Dev., 2016
Sonnenstuhl & Huth, Brain and Lang., 2002
Williams et al., ACL, 2020
Zaretsky & Lange, MaLT, 2016

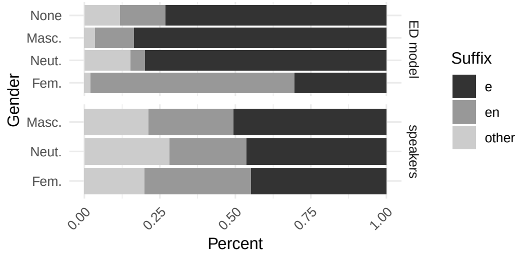

Figure 2: Plural class production by grammatical gender. Upper: Predictions from ED model trained without gender ("None") and with gender. Lower: 192 German speakers.
Model productions of /-e/ and /-(e)n/ follow the gender-conditioned distribution seen in Fig. 1, while speaker productions are much less sensitive to gender.

Suffix	Fixed effect	Est. β	SE	z	$Pr(>\|z\|)$
/-(e)n/	(Intercept)	-1.42	.22	-6.6	6e-11 ***
	gdr.masc	-1.02	.08	-13.5	2e-16 ***
	gdr.neut	-.36	.06	-5.9	5e-09 ***
	src.ED	-.13	.15	-.8	.40
	gdr.m:src.ED	-.74	.08	-9.9	2e-16 ***
	gdr.n:src.ED	-.27	.06	-4.4	1e-05 ***
/-e/	(Intercept)	.33	.17	1.9	.06 .
	gdr.masc	.34	.05	6.9	7e-12 ***
	gdr.neut	.58	.05	11.5	2e-16 ***
	src.ED	.47	.13	3.6	.001 ***
	gdr.m:src.ED	.38	.05	7.7	8e-15 ***
	gdr.n:src.ED	.40	.05	7.9	3e-15 ***

Table 2: Summary of fixed effects from logistic mixed-effect models (using the lme4 package in R; Bates et al., 2015), with random intercepts for participant and item, and sum-coded contrasts for gender (gdr) and data source (src).

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

Table 3: Correlations (Pearson's *r*, 95% CI in parens) between item-level production percentages for speakers and ED model *with* and *without* explicit grammatical gender indicated.
Correlation was evaluated across three bins per item and gender: percent /-e/ produced, percent /-(e)n/ produced, and percent all other productions. Item-level speaker data is more correlated to the productions of the model without gender ("None" in Fig. 2) than the model with gender.

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

Table 4: German plural suffixes with CELEX frequencies (Sonnenstuhl and Huth, 2002).

Linguistic addendum: German plurals

The German plural system comprises five main suffixes: /-(e)n/, /-e/, /-er/, /-s/, and /-∅/ (the “zero plural"). /-e/, /-er/, and /-∅/ can also combine with an umlaut over the root vowel; for simplicity, we focus only suffixes. /-e/ and /-(e)n/ are the two most frequent suffixes, in terms of both type and token frequency (Tab. 4). Grammatical gender is indicated on the article preceding the noun, as masculine *der*, feminine *die*, or neuter *das*. Gender is highly associated with plural class: most feminine nouns take /-(e)n/, while /-e/ and /-∅/ nouns are often masculine or neuter (Fig. 1). The phonological shape of a noun also influences its plural class; for example, most nouns ending with schwa take /-(e)n/.

Technical addendum: Encoder-decoder model

Neural encoder-decoder (ED) networks are a type of model which encodes an input sequence into a fixed vector representation and then incrementally decodes it into a corresponding output sequence. We use the Morphological Encoder-Decoder (Kann and Schütze 2016), a bidirectional recurrent neural network (RNN) architecture which has been proposed for cognitive modeling (Kirov and Cotterell, 2018). For the task of German number inflection, the ED takes as input the nominative singular 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). As a baseline, we trained a separate model with the same architecture on the same data without grammatical gender (e.g. W A H L). The encoder RNN incrementally processes each character, combining its input representation with the recurrent hidden state from the previous step; as it is bidirectional, this process runs both forward ($\langle f \rangle$ W A H L) and backward (L H A W $\langle f \rangle$) over the input string, and both representations are combined to produce the encoded vector. A separate decoder RNN incrementally produces the output string. At each time step, the decoder uses *attention weights* (Bahdanau et al., 2015) to recombine the encoded input, which it then combines with the recurrent hidden state from the previous step to predict the next output character. The model is trained to produce the noun's corresponding nominative plural form as output (e.g. W A H L E N).

For training, we used the 11,243 German nouns in UniMorph (Kirov et al., 2016; Fig. 1) as our corpus, and added noun gender data from another dataset (`https://github.com/gambolputty/german-nouns/`). We used 80% of the corpus for training, 10% for development (i.e. hyperparameter selection; based on dev set performance, we stopped training after 10 epochs), and 10% for testing (where the model achieved 88% accuracy). Following Corkery et al. (2019), we trained 25 separate random initializations of the same architecture, treating each model instance as a simulated “speaker". For evaluation, we provided each noun (Tab. 1) with each gender as input to each model instance, and aggregated the resulting productions (Fig. 2). Please see McCurdy et al. (2020) for further implementation details.