German Grammatical Gender Manages Nominal Entropy

Richard Futrell & Michael Ramscar
Stanford University
contact: futrell@stanford.edu

Introduction

Grammatical gender afflicts over half the world's languages, yet many hold it to be an arbitrary and redundant feature (Kilarski, 2007).

We propose a functional motivation for grammatical gender based on information theory. By **lowering the entropy of nouns in context**, grammatical gender **contributes to communicative efficiency**

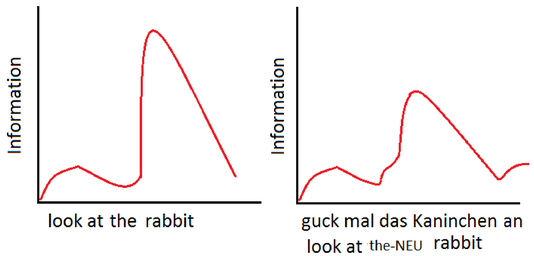

Figure 1. Hypothetical entropy rate without gender marking (English) and with it (German).

We test two consequences of this model in German. First, since Gender lowers the conditional entropy of nouns in the context *Definite article-Noun* (D-N), **German speakers should be able to use more informative nouns** in that context than English speakers.

Second, the pattern of gender assignment should reflect adaptation to the function of noun prediction.

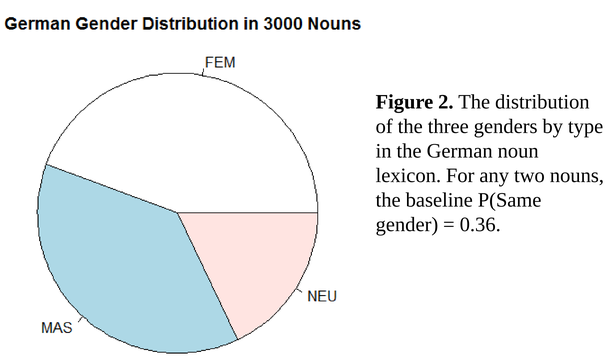

Figure 2. The distribution of the three genders by type in the German noun lexicon. For any two nouns, the baseline P(Same gender) = 0.36.

What does gender do?

We examined 3000 non-compound German noun types in the NEGRA II corpus (16,000 noun tokens). We used word frequencies after articles to calculate bare entropy and conditional entropy given gender. Then we calculated the entropy of English nouns in a sample of the same size in the NYT Gigaword corpus. The results:

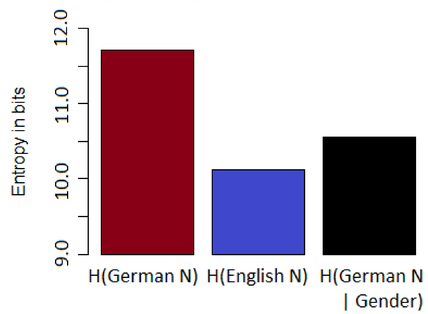

Figure 3. Entropy of German and English nouns after definite articles.

The entropy of English nouns is similar to the *conditional* entropy of German nouns. The higher entropy of bare German nouns could result from **increased lexical diversity in the context D-N**, confirmed below:

***Mean German Freq* = 2.12**
***Mean English Freq* = 4.93 (p < 0.001)**

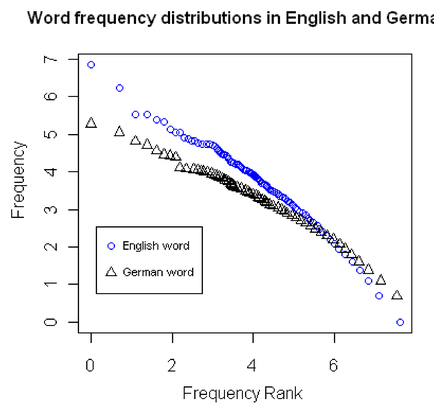

Figure 4. Frequency distribution of English and German lemmas in the context D-N.

How does it do it?

German gender sometimes groups semantically similar nouns (Zubin & Köpcke, 1986). This allows for two kinds of discrimination:

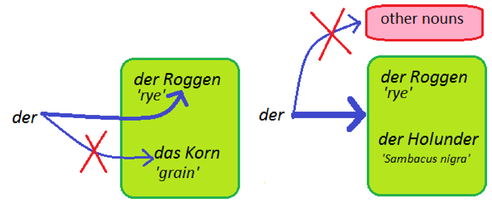

Figure 5. Gender may either discriminate between semantically similar nouns, or provide a cue for a semantic neighborhood.

We predict that **semantically similar nouns will have different gender when both nouns are frequent,** i.e. when speakers have to discriminate between them frequently.

Using the 3000 nouns in the CALLHOME German lexicon, we measured semantic similarity as distributional similarity in the Google 4-Gram corpus with gender information removed.

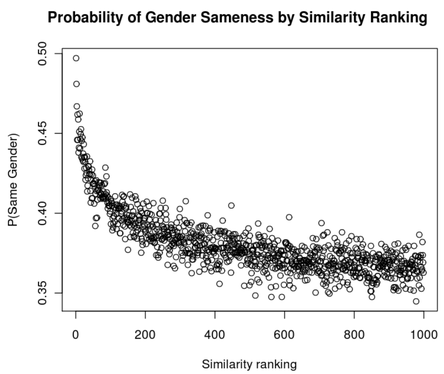

Figure 6. Probability that two nouns will share a gender based on their similarity ranking. The closest semantic neighbors (rank = 1) have just a 50% chance of sharing a gender.

Does a noun pair share a gender?	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.90	0.49	9.93	< 0.001
Log similarity	1.72	0.19	8.89	< 0.001
Log product frequency	-0.31	0.04	-8.88	< 0.001
Log similarity * Log product frequency	-0.10	0.01	-7.32	< 0.001

Table 1. Nouns are less likely to be in the same gender if one must frequently discriminate between them.

Next, we compared each noun to its 10 nearest distributional neighbors, and used logistic regression to predict whether the pairs in the neighborhood would share a gender (1 = same gender; 0 = different gender).

Conclusions

Gender emerges as a system for managing the rate of information transfer in language. Gender allows German speakers to encode more information into the channel without increasing demands on the hearer.

Gender lets German speakers deploy a greater variety of informative nouns in the context D-N. To use those same nouns, English speakers may make use of other predictive context, such as prenominal adjectives (see Ramscar & Futrell, 2011).

Further work should show how this function may or may not generalize to other gender-like systems, such as those manifested *after* nouns, and numeral classifiers.

Literature cited

Kilarski, M. (2007). On grammatical gender as an arbitrary and redundant category. In Douglas Kilbee, ed., History of Linguistics 2005: Selected papers from the 10th International Conference on the History of Language Sciences (ICHOLS X), John Benjamins, Amsterdam.

Zubin, D & Köpcke, K-M (1986). Gender and Folk Taxonomy: The Indexical Relation Between Grammatical and Lexical Categorization. In *Noun Classification and Categorization.* Philadelphia: Benjamins, North America, pp. 139–180.

We thank Melody Dye and Dan Jurafsky for help and advice.