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 (Kilaraki, 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.

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:

- German gender: 0.49
- English gender: 0.19

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, it allows German speakers to encode more information into the channel without increasing demands on the hearer.

Do a noun pair share a gender?
| Estimate  | Std. Error | p 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 |

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

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

- Semantic neighborhood (rank = 1) have just a 50% 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.
- 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.

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


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