Book Section:
Dye, M., Milin, P., Futrell, R. et al. (2017) A functional theory of gender paradigms. In: Kiefer, F., Blevins, J.P. and Bartos, H., (eds.) Perspectives on Morphological Structure: Data and Analyses. Brill, Leiden, pp. 212-239. ISBN: 9789004342910.

https://doi.org/10.1163/9789004342934_011

A Functional Theory of Gender Paradigms

Melody Dye [1], Petar Milin [2 3], Richard Futrell [4], & Michael Ramscar [2]

[1] Indiana University
[2] Eberhard Karls Universität Tübingen
[3] University of Sheffield
[4] Massachusetts Institute of Technology

Abstract

A central goal of typological research is to characterize linguistic features in terms of their functional role in a language. One longstanding puzzle for typologists concerns why certain languages employ grammatical gender, which assigns nouns to distinct classes. From a taxonomic perspective, gender specification can appear arbitrary, with little obvious correspondence between semantics and noun class. Gender has thus long been viewed as a useless ornament with no apparent rhyme or reason. However, there is an accumulating body of evidence that native speakers use determiners to guide lexical access. Here, we investigate whether an information theoretic perspective might shed some light on the communicative function of noun classification in German. We hypothesize that the system works to efficiently smooth information over discourse, making nouns both more predictable — and more equally predictable — in context. In line with these predictions, a large-scale corpus analysis reveals that German gender markers systematically reduce nominal entropy, facilitating the use of a more diverse (and more informative) set of nouns. Moreover, the structure of the gender system mirrors that of other subsystems of language, in that it provides systematic support for lower frequency forms. Thus, it is only from a taxonomic standpoint that gender's purpose can appear opaque: Our findings indicate that German gender classes conform to a tight 'discriminative' logic, employing a structure system of semantic clusters and contracts to facilitate lexical processing.

> "In German… every noun has a gender, and there is no sense or system in the distribution; so the gender of each must be learned separately and by heart. There is no other way. To do this one has to have a memory like a memorandum-book. In German, a young lady has no sex, while a turnip has. Think what overwrought reverence that shows for the turnip, and what callous disrespect for the girl…:
>
> > **Gretchen**: Wilhelm, where is the turnip?
> > **Wilhelm**: She has gone to the kitchen.
> > **Gretchen**: Where is the accomplished and beautiful English maiden?
> > **Wilhelm**: It has gone to the opera.
>
> … a tree is male, its buds are female, its leaves are neuter; horses are sexless, dogs are male, cats are female – tomcats included, of course; a person's mouth, neck, bosom, elbows, fingers, nails, feet, and body are of the male sex, and his head is male or neuter according to the word selected to signify it, and **not** according to the sex of the individual who wears it – for in Germany all the women wear either male heads or sexless ones; a person's nose, lips, shoulders, breast, hands, and toes are of the female sex; and his hair, ears, eyes, chin, legs, knees, heart, and conscience haven't any sex at all. The inventor of the language probably got what he knew about a conscience from hearsay."

Mark Twain, (1880) "*The Awful German Language"*

> "The confusions that occupy us arise when language is like an engine idling, not when it is doing work."

Wittgenstein (1953), *Philosophische Untersuchungen*

1 Introduction

In his humorous account of the "awful" German language, Mark Twain draws attention to a puzzle posed by many of the world's languages: grammatical gender. As often as not, the languages of the world assign objects into seemingly arbitrary (and often seemingly sexist) noun classes that lack any transparent purpose (Corbett 1991). Historically, this led some scholars to conclude that grammatical gender is senseless: William of Ockham considered gender to be a meaningless, unnecessary aspect of language, an obvious candidate for his famous razor; Baudouin de Courtenay described gender as a deformity, an unfortunate historical accident that was responsible for a range of human afflictions, including nightmares, pathological behavior, erotic and religious delusions, and sadism (Kilarski, 2007). Few other linguists have held noun class to be responsible for all of the world's ills; but few have warmed to its virtues either. The consensus is neatly summarized by Leonard Bloomfield (1933): "[t]here seems to be no practical criterion by which the gender of a noun in German, French, or Latin [can] be determined."

Not only have gender systems been branded as meaningless, but they are fiendishly difficult for non-native speakers to learn, a state of affairs that prompted the developmental psychologist Michael Maratsos (1979) to conclude:

> "The presence of such systems [German gender] in a human cognitive system constitutes by itself excellent testimony to the occasional nonsensibleness of the species. Not only was this system devised by humans, but generation after generation of children peaceably relearns it."

While many linguists have reconciled themselves to the idea that gender has evolved its negative consequences for no reason, Charles Darwin was less sanguine about such matters: "The sight of a feather in a peacock's tail, whenever I gaze at it, makes me *sick*," he famously wrote.[1] In the 1800's, Darwin's pursuit of evolutionary explanations for such apparent anomalies revolutionized our understanding of biology. Indeed, his ruminations on the peacock's tail helped develop the theory of sexual selection: Darwin hypothesized that while the extravagance of the male peacock's train might prove hazardous to its health, females would often opt for mates with more ornate plumage, leading to reproductive success for showier males. Hence even the seemingly 'absurd' and risky feather display of a male peacock might still have an adaptive purpose. In this chapter, we adopt Darwin's stance in analyzing the place of grammatical gender in German, seeking to elucidate a functional role for gender marking in facilitating communicative efficiency.

1.1 Some Proposed Functions of Noun Class

Grammatical gender is an obligatory morphological system found in many languages that groups nouns into a small number of mutually exclusive classes, and marks neighboring words (such as articles and adjectives) for agreement. Many languages, such as French and Spanish, divide nouns into two distinct classes: masculine and feminine. Others, like German and Russian, add a third neuter category, yet even more are possible; Swahili has six (Corbett, 1991). Speaking broadly, a noun's gender specification tends to be semantically arbitrary, with little obvious correspondence between the conceptual properties of the referent and its noun class, and substantial

[1] Letter 2743 – Darwin, C. R. to Gray, Asa, 3 Apr (1860), Darwin Correspondence Project

cross-linguistic variation (Vigliocco et al., 2005).

While not all researchers consider noun class to be purely ornamental, many of the functions that have been proposed for other languages have only limited applicability in German, the focus of the present study. One hypothesis is that gender marking assists comprehension processes by linking temporally separated elements in discourse, establishing local and global coherence. For instance, in some languages – but not German – agreeing gender markers can facilitate freedom in word order by marking which words describe the same thing. As can be seen in (1), Latin 'attributive' adjectives need not appear in a fixed position relative to nouns; since suffixes are declined for gender, case, and number, it is clear when an adjective and a noun belong together:

(1) ultim-a Cumae-i ven-it iam carmin-is aet-as
last-NOM.FEM Cumai-GEN.NEU came now song-GEN.NEU last-NOM.FEM
"The last age of the Cumaean song has now arrived."[2]

Yet German does not have this functionality. Only attributive adjectives are marked for agreement, and those adjectives cannot appear anywhere other than immediately before a noun:

(2) die *große* Frau sah das Kind
(3) *die Frau sah das Kind *große*

Perhaps the most concrete suggestion that has been put forward for German gender's function, is that agreement between gender markers and anaphoric pronouns facilitates reference tracking (Zubin & Köpcke, 1986; Koval, 1979; Heath, 1975, *i.a.*). Consider the following:

(4) der Krug fiel in die Schale, aber er zerbrach nicht
the.MAS jug fell into the bowl.FEM but it.MAS broke not
'The jug fell into the bowl, but it (the jug) didn't break.'

In this instance, the referent of the pronoun 'it' is unambiguous, because 'it' must have a MASCULINE referent (which in this case must be the jug, not the bowl). However, even this proposal suffers shortcomings. For one, the existence of semantic regularities in noun class works *against* reference tracking, by increasing the probability that confusable nouns will be referenced with the same gendered pronoun (Lakoff, 1986). For another, German grammar frequently does not permit its speakers to rely on gender for this kind of discrimination (Claudi, 1985).

As these examples illustrate, gender may play different roles in different languages. Indeed, there is a growing body of evidence attesting to cross-linguistic differences in morphosyntactic processing, showing substantial variation in how listeners make use of gendered determiners in discourse (see e.g., Miozzo & Caramazza, 1999; Schriefers & Teruel, 2000). Accordingly, it would be a mistake to treat all systems called "noun class" as the same thing and to ignore the details of how, when, and where language speakers mark gender (see also MacWhinney, Bates, & Kliegel, 1984).

[2] From Virgil's Aeneid, cited in Matthews 1981 and Evans 2010.

In determining the function of noun class in a given language, it is critical to examine the part that gender marking plays both in communication between current speakers (information processing) and in transmission between generations (learning). In what follows, we conduct precisely such an examination from the vantage point of information theory. While information theory is typically considered in the context of modern computing and engineering, it provides a useful lens through which to consider human language. In particular, its mathematical toolkit offers a precise means of quantifying how information is distributed across a language. By measuring systematic variations in that distribution in German, we are able to investigate how gendered determiners aid efficiency in linguistic processing. The findings we present here provide compelling support for the idea that grammatical gender is no mere ornament. On the contrary, gender appears to be an invaluable resource for regulating the flow of information between speakers.

2 An Information Theoretic Approach

To understand the import of information theory to this problem, it is useful to contrast the lens it offers against that of the standard linguistic model. Since antiquity, language has mainly been conceived of in terms of a single, dominant metaphor: that of the direct material exchange of messages. According to the 'conduit' of metaphor of communication, a speaker packs the content of a message into words, which a listener unpacks at the other end. Utterances are supposed to somehow 'contain' their meanings, much as a stamped envelope contains a letter (Reddy, 1979). This metaphor for understanding language is pervasive in folk psychology, and is reflected in a broad array of psychological and linguistic theories (Sperber & Wilson, 1996).

Yet the conduit metaphor is neither inevitable nor irresistible. Conveyance systems, such as the mail or the carrier pigeon, are not the only means by which human societies have communicated at a distance, and an indirect alternative, in which messages are telegraphed across space and time, rather than physically conveyed by transport, has long been available. In telegraphy, no material copy of the message is ever sent. Instead, the message is translated into a physical signal that can travel the distance required. Successful communication relies on both the sender and receiver sharing the same code, such that the receiver can discriminate the original message from the received signal (Holzmann & Pehrson 1994). Modern digital communications are conceived of within precisely this framework, which Shannon (1948) formalized in information theory.

These two models of communication offer radically different lenses through which human language can be seen. On the direct transfer model, communication is at once deterministic (in the sense that words, like physical packages, are assumed to convey a certain determinate content), and singular (in the sense that any given communicative exchange is an isolated event, independent of the broader communicative context, or the prior history of the words or their speakers; Campbell, 1982). By contrast, on the indirect signaling account, communication is predictive and probabilistic. No communicative act occurs in a vacuum; rather, it occurs within the context of a larger linguistic system, governed by extensive, quantifiable regularities. The likelihood of any given message can only be assessed against the distribution of other possible messages that might have been selected instead.

Whereas problems with the first model are well attested (for notable criticisms, see e.g., Wittgenstein, 1953; Quine, 1951; Ramscar & Port, 2015; Baayen & Ramscar,

2015), there is much to recommend the second, particularly as formalized in information theory. Like artificial communication systems, natural languages involve a sender and receiver, a code, and a basic transmission problem. Moreover, they too are indirect means of information exchange (Mandelbrot, 1953; Ramscar & Baayen, 2013). From this perspective, human languages can be seen as complex systems that have evolved over thousands of years and billions of speakers to optimize information flow in communication, and to balance the countervailing demands of learnability and fluent processing (see also Blevins, Milin, & Ramscar, *this volume*).

2.1 The Discrimination Problem

> "…consider a coding scheme devised to transmit four experiences: the experience of a fountain, the experience of a fountain pen, the experience of an orange, and the experience of orange juice. Assume a code, shared by encoder and decoder, specifying that the four experiences are signalled by the digit strings 00, 01, 10, and 11, respectively. When seeking to communicate the experience of a fountain pen, the speaker will encode 01, and thanks to the shared code, the listener will decode 01 into the experience of a fountain pen. There is no need whatsoever to consider whether the individual ones and zeroes compositionally contribute to the experiences transmitted. Thus, we can view language […] as a signal that serves to discriminate complex experiences of the world."
>
> Baayen & Ramscar (2015)

Within the discriminative framework shared by learning and information theory, language is best described as a probabilistic enterprise in which speakers and listeners cooperate in order to discriminate the content of an intended message from possible alternatives. Formally, the process can be characterized as one of iterative uncertainty reduction: Just as each forking branch in a decision tree further delimits the space of final outcomes, so each utterance (or articulatory gesture) further narrows the range of possible messages (Ramscar & Baayen, 2013). In assessing the dynamics of this process, it is possible to identify both the uncertainty at a given point, and the extent to which it is subsequently reduced. For instance, in context, a speaker's choices can be seen as more or less constrained, corresponding to more or less uncertainty about which word will be uttered next. The more freedom the speaker has in selecting amongst alternatives, the greater the uncertainty, and correspondingly, the more difficult the discrimination problem.

Taking a discriminative approach to communication lays bare the difficulties that nouns pose for language users. In most languages, nouns (both common and proper) are the most *diverse* part of speech, meaning that in any instance in which a noun occurs, the number of other possible alternatives is at its highest, and the discrimination problem is at its peak. This is supported by numerous findings on speech errors. For instance, one of the most common places disfluencies are likely to occur in English is at the determiner preceding a noun; and the more complex the noun is, the more likely a disfluency (Clark & Wasow, 1998). Similarly, nouns are the most common sites for incorrect lexical retrieval and a host of other processing problems (Vigliocco, 1997).

Critically, for our purposes, difficulties such as these have been shown to correlate with the information-theoretic measure of *entropy*, a measure that can be used to quantify the uncertainty over which word will appear in a given context. Entropy offers a particularly useful compression scheme for conceptualizing linguistic uncertainty. While the predictability of a card draw or coin flip is easy to grasp, uncertainty is more difficult to intuit when possible outcomes are numerous, or sequentially dependent, or

where probabilities are varied, as is the case for lexical distributions, which comprise thousands of words of widely varying frequencies. Entropy helpfully collapses a multi-dimensional construct down to a single point on a continuum.

Formally, the entropy *H* over such a distribution is a measure of the expected value of information ('surprisal') over the full range of lexical items (Shannon, 1948):

$$H(p) = H(X) = -\sum_{x \epsilon X} p(x) log_2 p(x) \quad (1)$$

When comparing across similarly shaped distributions, entropy will tend to rise as the number of possible outcomes grows. This means that for languages such as English and German, in which the number of noun types outstrips other part-of-speech categories, speakers of both languages will be confronted with most uncertainty when the next item in a sequence is a *noun*. Thus, in example (5), the entropy of possible noun continuations (marked *!*) will be higher than for possible verb continuations (marked *#*).

(5) Yesterday I *#* visited the *!* doctor.

Fortunately, speakers have various resources at their disposal for making a particular lexical choice more or less predictable in context. One possibility is to rely on the preceding discourse as a form of scaffolding. For instance, compared to the sparse semantic context provided by (5), the noun *doctor* is far more predictable following the comparatively rich context provided by (6):

(6) Yesterday when I went to the hospital I visited the *! doctor*.

Noun class offers an efficient, systematic way of implementing the same principle. Consider the German equivalent of (5) in (7):

(7) *Gestern besuchte ich den ! Arzt*
yesterday visited I the.MASCULINE *!* doctor

While the context is the same as in (5), the uncertainty about the following noun in (7) is greatly reduced by comparison. The following noun must belong to the MASCULINE noun class, and thus nouns of all other genders are eliminated as possible candidates in this context. In short, by systematically partitioning nouns into different classes, a gender marker effectively prunes the space of subsequent possibility, delimiting the set of upcoming nouns to class-consistent possibilities.

There is an accumulating body of evidence that gendered articles guide lexical prediction in precisely this way. Among native speakers of gendered languages, a variety of experimental paradigms, including naming times (Schriefers, 1993), lexical decision (Grosjean et al., 1994), word repetition (Bates et al., 1996), artificial grammar learning (Arnon & Ramscar, 2012), and ERP (Van Berkum et al., 2005; Wicha, Moreno, & Kutas, 2004) have shown that gender facilitates processing when a marker is consistent with a following noun, and inhibits it where there is a mismatch. Auditory gating studies have proved particularly revealing. In such tasks, subjects encounter a word fragment within a clipped auditory sequence, and are asked to produce the target

word. When gender information is provided, French subjects correctly identify the target at shorter durations, and with greater confidence. Moreover, gender information not only significantly reduces misidentifications, both in terms of types and tokens, but also limits errors to gender-consistent candidates (Grosjean et al., 1994). In a similar vein, in tip-of-the-tongue (TOT) states, Italian subjects can reliably guess the gender of the noun they are trying to retrieve, even when they cannot produce it (Vigliocco, Antonini, & Garrett, 1997).

These findings are paralleled in studies of visual search. In a study of French speakers, Dahan et al. (2000) asked subjects to view a visual display with a variety of possible referents, while they listened to instructions such as *Cliquez sur le bouton* [*Click on the.MASC button*]. When gender information was provided at the determiner, listeners rapidly shifted their attention to gender-consistent referents, ignoring potential phonological competitors. Lew-Williams and Fernald (2007) report a comparable result for Spanish-speakers, finding that both children and adults are faster to orient to the correct referent on trials when nouns of different genders are displayed than on trials showing nouns of the same gender. Taken together, these results support the conclusion that gendered articles facilitate processing by restricting the space of subsequent possibility.

2.3 Managing and Redistributing Entropy

In understanding the function of gender from this perspective, it is critical to note that gender does not reduce overall entropy, so much as *redistribute* (or manage) it—increasing the entropy of articles, while decreasing the entropy of the nouns that follow them. From a processing perspective, this is consistent with Zipf's famous 'Principle of Least Effort,' which holds that human behavior is shaped by a bias to minimize people's "average rate of work-expenditure over time" (Zipf, 1935; 1949).

On Zipf's loosely psychological account, communicators seek to balance efficiency on the one hand, and comprehensibility on the other, and these opposing forces minimize communicative effort over time. For example, in a lexicon in which each distinct meaning was assigned a separate word, there would be zero ambiguity, but at a significant processing cost to the speaker engaged in word retrieval. Conversely, a vocabulary comprising a single word would be maximally efficient for the speaker, but "represent the acme of verbal labor" (21) for a listener. Zipf argued that language's characteristic statistical structure reflects a compromise that balances the desire for a many-to-one code (in which there is a single, maximally frequent word) against the desire for one-to-one code (in which there are a vast number of low-frequency words). In the terms of optimal coding theory, these balancing forces of unification and diversification can be framed as a compromise between 'word-by-word' coding and 'large-block' coding (Mandelbrot, 1953). Thus, the problem of language design is one of how to distribute the information necessary to discriminate the repertoire of possible messages across acoustic signals (Baayen & Ramscar, 2015).

Once that has been established, the most efficient means of transmitting information across a channel is at a constant rate at (or approaching) the channel's capacity (Shannon, 1948). Indeed, a raft of empirical findings suggest that in accordance with this principle, speakers distribute uncertainty evenly across discourse, in both text and speech. One prediction that comes out of this, is that if the sentences of a given text are equally informative when encountered in context, this is only because the meaning constructed from earlier parts has generated an informative context that reduces the

entropy of later parts. In the limit, this suggests that when this contextual scaffolding is stripped away, utterances should become increasingly informative the deeper embedded in discourse they are. This basic growth pattern has been demonstrated empirically: In a classic study of articles in the Wall Street Journal, Genzel and Charniak (2002) found that local sentence entropy increases as a function of sentence number, an effect that is driven both by which words are used and how the words are used (i.e., both lexical and syntactic causes). The effect has since been replicated across languages and genres (see also Genzel & Charniak, 2003; Keller, 2004; Qian & Jaeger, 2009).

At the same time, a growing body of evidence supports the idea that in language use, people deftly manage the rate at which information is encoded in linguistic signals, avoiding excessive peaks and troughs in entropy across messages (Aylett & Turk, 2004; Levy, 2008; Jaeger, 2010). One domain in which this has been rigorously tested is speech production, where speakers have been found to smooth information over the acoustic signal by systematically modulating the signal's properties. Varying acoustic duration is one way to accomplish this: articulating unpredictable segments more slowly than predictable ones, and shortening, undershooting, or omitting highly predictable segments (see Gahl, 2012 for a review). These predictions have been substantiated in multiple studies. For instance, Aylett and Turk (2004) found that an inverse relation obtains between a syllable's duration and its predictability in context. Comparable findings on informativity and articulatory effort have been made for words (Bell et al., 2009), morphemes (Pluymaekers, Ernestus, & Baayen, 2005), consonants (Van Son & Van Senten, 2005), and multi-word sequences (Gahl & Garnsey, 2004; Kuperman & Bresnan, 2012). Durational effects have even been replicated in typing (Priva, 2010).

Similarly, in anticipating upcoming words that are information rich, speakers may pause or otherwise delay (Goldman-Eisler, 1958). Predictability also affects specific lexical choices in spontaneous speech and reading aloud: When what they are about to say is predictable, speakers are more likely to employ contractions (Frank & Jaeger, 2008), to omit optional function words (Jaeger, 2010), to use a pronoun referent instead of a full noun-phrase (Tily & Piantadosi, 2009), and to produce fewer disfluencies (Tily et al., 2009). Conversely, when speakers repeat or mimic syntactic constructions in discourse, they temper syntactic redundancy with the selection of more informative, less predictable words (Temperley & Gildea, 2015).

Parallel investigations have been carried out cross-linguistically, with promising results. In a large-scale corpus study spanning eleven Indo-European languages, Piantadosi et al. (2011) found that a word's length is better captured by its average predictability in context than by its raw frequency, with more informative words taking longer forms (see also Manin, 2006). Likewise, in a cross-linguistic comparison of reading aloud data, Pellegrino, Coupé, and Marsico (2011) report that while the various languages under study achieve roughly comparable rates of information transfer overall, they strike markedly different balances between information density and speech rate in doing so: in languages with less information per syllable, syllables tend to be spoken faster, and vice versa.

These findings make clear that language distributions (and speakers) ably regulate the uncertainty associated with temporal dynamics and lexical choices. Gender markers may simply serve as another resource by which to accomplish this: If gendered articles serve to redistribute nominal entropy, this will smooth potential spikes in information, helping speakers maintain a more constant entropy rate.

3 Noun Class and Entropy Reduction in German

Our proposal is that noun class systematically narrows the set of candidates that follow a gender marker, thereby reducing the amount of information that a noun would convey on its own. As a first test of this hypothesis, we conducted an analysis of nominal entropy distributions in the German mega-corpus Stuttgart deWaC.[3] German is a language with a binary number system (singular and plural), three-class gender system (masculine, feminine, and neuter), and four grammatical cases in which nouns can occur (nominative, accusative, dative, and genitive). Accordingly, to assess the influence of gender marking on nominal entropy, the entropy of all the nouns within each case (2) was compared to the conditional entropy of those nouns following articles marked for gender and number (3).

$$\begin{aligned} H(N) &= -\sum_i P(N_i) log_2 P(N_i) \\ &\approx -\sum_i \frac{\text{Count}(N_i)}{\text{Total}(N)} log_2 \frac{\text{Count}(N_i)}{\text{Total}(N)} \end{aligned} \tag{2}$$

For instance, for nouns following the masculine nominative article *der*, the conditional entropy would be given by:

$$\begin{aligned} H(N) &= -\sum_i P(N_i|der) log_2 P(N_i|der) \\ &\approx -\sum_i \frac{\text{Count}(N_i|der)}{\text{Total}(N|der)} log_2 \frac{\text{Count}(N_i|der)}{\text{Total}(N|der)} \end{aligned} \tag{3}$$

Consistent with our suggestion that German gender serves to reduce uncertainty about upcoming nouns in discourse, we found:

$$\begin{aligned} \text{Entropy (nouns)} &> \text{Entropy (nouns|definite article)} \\ &> \text{Entropy (nouns|definite article, case)} \\ &> \text{Entropy (nouns|definite article, case, gender)} \end{aligned}$$

[3] The SdeWaC is a subset of the WaCky corpus, which comprises more than 44 M sentences, 850M word tokens, and 1.1 M word types (Faaß & Eckart, 2013; Baroni, Bernardini, Ferraresi, & Zanchetta, 2009). The corpus was first annotated with fine-grained part-of-speech categories using the RFTagger (Schmid & Laws, 2008), and article contractions were expanded (*im -> in dem*). Every noun that immediately followed a definite article was extracted with its gender, case, and number tags, and tabulated.

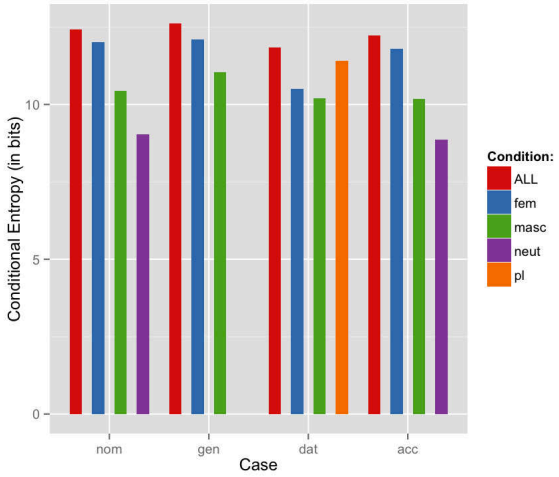

Fig 1. Noun entropy conditioned on case and number, irrespective of gender vs. gender sensitive. Notice that because of syncretism, not every category is represented independently for each case; German lacks any morphological distinction between feminine and plural articles in the nominative, accusative, and genitive cases, and between masculine and neuter articles in the dative and genitive cases. In this analysis, forms that took the same marker within a given case were tabulated together (e.g., for nominative, both feminine and plural nouns contribute to the entropy calculation for 'die').

These results show that, as expected, in each of the German cases, gender markers significantly reduce nominal entropy (Figure 1; The same qualitative results were obtained in an analysis of the Negra II corpus of German newspapers, Skut et al. 1997).

To further test this hypothesis, we then examined the effect of noun class marking on the distribution of nouns in German. By effectively partitioning the noun space, gender markers should offload some of the uncertainty about the upcoming noun onto the determiner, thereby smoothing entropy over the marker-noun pairing. Accordingly, when prenominal class marking is present, the following noun should be relatively well-predicted, compared to cases in which its class goes unmarked. Assuming that communicators aim to keep uncertainty relatively constant, and that gender marking offers an effective means of selectively modulating uncertainty, German speakers should make use of a greater variety of nouns when noun class marking is present than when it is absent.

The German plural offers an illustrative test case. While all German singular nouns are marked for gender, plural nouns are not. Accordingly, following a definite article, speakers should employ a more diverse (and more informative) set of nouns in the singular than in the plural. A measure of the difference in the overall lexical diversity of the two noun types in this context can be estimated by calculating their type/token ratio (while holding the sample size constant), with a higher type/token ratio suggesting a greater diversity of nominal usage. Conveniently, this metric is simply the inverse of average frequency, allowing for a straightforward test of this hypothesis: the lower the average frequency, the greater the diversity of nominal usage.

Consistent with our hypothesis, an examination of Determiner-Noun contexts in the SdeWaC revealed that the singular nouns in our sample had a higher type/token ratio than the plural nouns. When the frequencies for singular and plural nouns are

normalized to per-million occurrences, the mean frequency for singular nouns is 0.75, and that of plurals is 1.43; correspondingly, lexical diversity for singular and plural nouns is 1.33 and 0.70, respectively. German plurals, which are not gender-marked, thus show a substantial reduction in lexical diversity relative to singulars, indicating that gender catalyzes the use of a wider array of nominal forms.

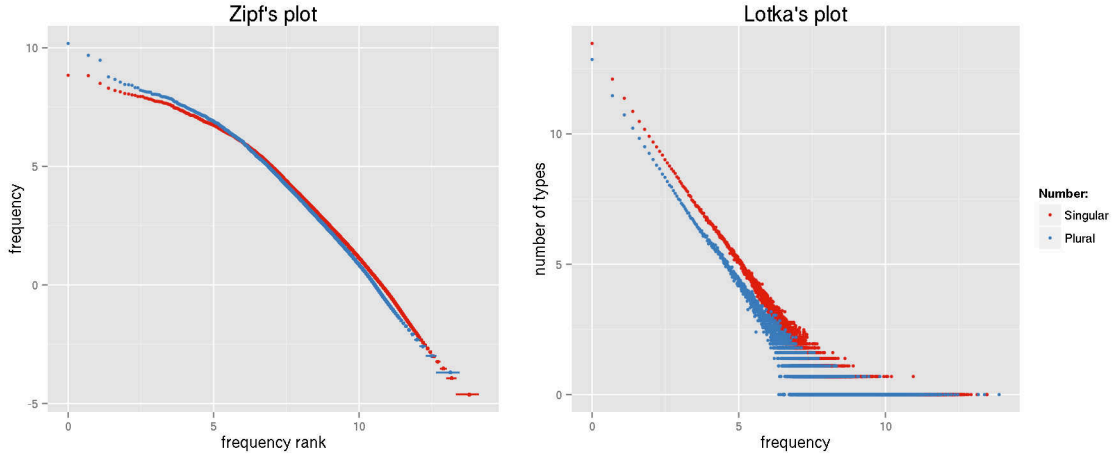

Fig 2. The frequency distributions of German singular and plural nouns following a determiner. These distributions are plotted in two complementary ways: While the Zipf-plot (left panel) plots frequency rank by frequency, the Lotka-plot (right panel) plots frequency by number of different word types; both are shown in a log-log plane. In a sense, the plots are showing each other's tails (c.f., Chen & Leimkuhler, 1986; Kunz, 1987).

Figure 2 shows the distributional impact of gender-marking. The figure represents the (extremely) skewed frequency distributions for singular (gender-marked) and plural (unmarked) nouns following determiners. The Lotka-plot indicates that whereas there are significantly more singular noun types with low frequencies, the inverse is true for plurals, which cover a wider range of the most frequent types. The Zipf-plot echoes this trend, revealing that the difference in nominal frequencies is most pronounced in the high frequency range; within that uppermost band, a singular noun of a given frequency rank will (on average) be markedly lower in frequency than its plural equivalent.

Interestingly, when determiners are treated as mere case markers, independent of gender and number, and their following distributions are analyzed separately, the lexical diversity of following nouns is *equal*, on average. This is, again, consistent with the suggestion that languages (and hence speakers) are finely attuned to the uncertainty of their productions, exploiting the varied resources at their disposal to keep entropy smoothed.

3.1 Semantics and the function of noun class

One question that arises is whether this partitioning of nouns into classes is arbitrary, or whether there might be a hidden logic behind it. Recall the damning words of Mark Twain—why the sexless young maiden, the female tomcat? Is there really no sense or sensibility to gender? In order to understand how noun class might best be configured to facilitate communication, it is important to consider the wider functional implications of uncertainty reduction in language use.

Recall that under the standard metaphor, language is conceptualized as a process of encoding, transmitting and decoding of tokens of meaning types. These meaning types have been assumed to be taxonomically organized, and encoded and decoded by rules that allow messages to be generated from them. From this perspective, the challenge facing both language learners and theoretical linguists is inductive: the correct taxonomy of meaning types and generative rules for a given language must be inferred from whatever data is available to the learner or theorist.

However, in contrast to inductive models, both information theory and empirically grounded psychological theories of learning describe deductive processes based on prediction and discrimination (Ramscar & Baayen, 2013; see Shannon, 1948; Kullback & Leibler, 1951; Rescorla & Wagner, 1972). Information theory sees "the fundamental problem of communication [as] that of reproducing at one point, either exactly or approximately, a message selected at another point" (Shannon, 1948). Seen from this perspective, communication need not consist in the transmission of tokens corresponding to fixed semantic types between speakers. Rather, it can be seen as a process in which a speaker reduces a hearer's uncertainty about the meaning of a message by whatever means are available (Ramscar et al, 2010).

For example, when a German speaker uses the expression *der Hund* "the dog", the masculine-gendered article *der* helps the hearer to expect a dog as the referent. From this perspective, not only the token *Hund* but also the gendered article *der*—and indeed the entirety of the surrounding predictive context, including any verbs and adjectives—helps the hearer to form the belief that the speaker wants to say something about a "dog". Linguistic communication can thus be seen as a probabilistic process in which a speaker helps a listener to predict, either exactly or (more often) approximately, the speaker's intentions.

These very different models of the way that language works yield very different predictions about the function of noun class. The taxonomic approach leads naturally to the prediction that noun class adds to the taxonomy of meaningful words in a given language. For example, Lakoff's (1986) 'Domain of Experience Principle' holds that nouns that occur in similar contexts tend to have the same gender. Gender is seen as mapping to an abstract "semantic field" for the purposes of transmitting meaning. However, since the task of identifying exactly what these semantic fields actually *are* has proven to be difficult in many languages, this idea is often augmented by the supplementary assumption that although many semantic fields have a dominant gender, each of these comes with a set of individually specified exceptions (Zubin & Kopcke 2007, 1981; Zubin 1992). That is, from the taxonomic perspective, noun classes tend to map to a semantic field, except in all cases where they *don't*.

However, a system in which semantic regularities are not immediately straightforward might actually *benefit* discrimination. The benefits of such dispersion are noted by Zubin and Köpcke (1986), who show that while nouns in the semantic class of 'kitchen implements' in German are evenly distributed among the genders without any obvious or sensible pattern, the "patterning" of their dispersal may actually facilitate reference tracking among objects. In short, the suggestion is that when nouns that occur in similar contexts are assigned different genders, this facilitates discrimination between possible referents.

On the other hand, Twain may have been exaggerating slightly; there do appear to be semantic patterns to gender assignment in German, though they are rife with exceptions (Zubin & Köpcke, 1986; Lakoff, 1986). The existence of such semantic regularities makes clear that noun classes are not distributed so that *all* similar nouns

receive different genders. (In the limit, such a scheme might prove redundant from the point of view of maintaining a constant entropy rate.) Instead, the German gender system may be optimized in a different way. For example, almost all German alcoholic drinks are in the masculine class, except *beer*, which is neuter. This state of affairs is mirrored for non-alcoholic beverages, which also tend to be in the masculine class, with one notable exception—*water*, which is also neuter (as are the common words for drink and beverage). Taxonomically, a grouping of gin and juice and coffee on the one hand, and beer and water on the other, makes little sense. Yet a class division between the drinks that might be more or less expected in a given context does, because a deductive discriminative process works by eliminating possible interpretations that are *not* intended (Shannon, 1948; Rescorla & Wagner, 1972; Ramscar et al, 2010).

To flesh out this idea, compare the information requirements for helping someone predict that Beethoven rather than Mozart will be the topic of a sentence, as compared to helping someone predict that it will be Villa Lobos rather than Schoenberg (lesser known 20th Century composers). In discourse about composers of classical music, both Beethoven and Mozart— by dint of their fame and presence in any educated Westerner's general knowledge—will be highly predictable in context. Thus, while much could be gained from deploying a contextual cue that eliminates Mozart as a possible topic as opposed to Beethoven (or vice versa), little could be gained from eliminating the relatively obscure and unpredictable Villa Lobos or Schoenberg. Contextual cues that are specifically informative about high frequency items will be very useful for discriminating between those items.

Even if the topic of discourse does turn out to be Villa Lobos or Schoenberg, cues that favor the elimination of highly predictable competitors will still be more valuable than cues that favor the specific prediction of one over the other. Because Mozart and Beethoven will be strongly expected candidates in discourse about composers, a cue that eliminated one or both of them from consideration would be a boon for communicative clarity, as it would improve the predictability of *both* Villa Lobos and Schoenberg. This is not to say that contextual information that discriminates Villa Lobos from Schoenberg might not also be helpful here, but rather that that information will only be relevant *after* competition from Mozart and Beethoven has been reduced or eliminated.

As this example illustrates, depending on the distribution of items in a semantic class, both semantic clustering and semantic dispersal could be employed to optimize the use of gender information for discriminating between alternatives of differing probabilities. For example, to assist with overall entropy reduction, a noun class system might fruitfully assign Beethoven and Mozart to their own classes, while grouping Villa Lobos and Schoenberg together in another. Indeed, in terms of informativity, it might be perfectly sensible if Villa Lobos and Schoenberg were classed alongside more obscure composers from other classical periods, even if this makes relatively little sense taxonomically. This logic can begin to help explain why German puts what are historically its most common drinks—beer and water—in a class apart from most other beverages.

3.2 Testing Semantics

The notion that German noun class is informative is compatible with both a taxonomic and a discriminatory approach to language. To the extent that the two approaches make different predictions, the differences are in the details. While both models predict a

correlation between semantics and gender, the taxonomic model leaves the exact nature of that relation opaque and filled with exceptions. By contrast, the discriminatory model suggests that 'exceptions' are likely nothing of the sort, reflecting instead the properties of the underlying system. The discriminatory model thus makes an intriguing prediction: not only should we expect to find a positive correlation between semantics and noun class, but we should also be able to detect systematic patterns where semantics and noun class diverge.

For a gender system to be maximally functional, it needs to reduce the uncertainty of an upcoming noun in context by *narrowing the search space* of likely candidates. That is, it needs to discriminate against alternative nouns on the basis of their likelihood. To ideally meet this requirement, such a system should assign different genders to nouns that are both semantically similar and potentially highly confusable in context. But this raises an intriguing question: how might this be achieved? In practice, semantic considerations at the local and discourse level will have significantly altered the shape of the likelihood distribution, making some nouns far more likely in context, and others considerably less so. However, absent a means of making entropy reduction semantically interpretable – a task that is beyond the scope of the current work – this observation is less than illuminating.

To try and shed some light on this question, consider the following possibilities:

1.) Noun class might discriminate semantically similar nouns that co-occur together regularly, such as gin versus tonic, or coffee versus tea.

2.) Noun class might discriminate semantically similar nouns that differ by frequency, such as water (high-frequency) versus root beer (low-frequency).

3.) Noun class might discriminate nouns that are highly likely in a certain context, but which are semantically distinct, such as drinks versus food.

As we noted in our composers example above, the degree to which one strategy or another is most appropriate for a given noun will depend both upon its overall likelihood, and the degree to which it is already predicted when a gender marker occurs. Thus 1) will work better when there is a higher degree of certainty about the specific noun that will occur, whereas 3) will be a better fit when there is a lower degree of certainty. The degree of uncertainty will always depend on the specifics of the particular noun: its frequency, the frequency of its neighbors, and the contexts in which it (and its neighbors) are encountered. An optimal system should tailor its level of support for each noun based on these factors, supplying different information depending both upon the overall likelihood of the noun, and the likelihood of there being other discriminatory information available in context.

To gain a better understanding of how this might function in German, we examined the fine-grained relationship between semantics, contextual confusability, and noun class. Specifically, using a Generalized Additive Model (GAM) with binomial link-function (*mgcv* package in R Statistical Computing Environment; see Wood, 2006; 2011; R Core Team, 2015), we attempted to predict gender sameness for pairs of nouns based on the pair's frequency, pointwise mutual information (PMI), and semantic

similarity.[4] Our modeling results were validated with a bootstrap sampling technique (N=1000 simulation runs). To remain conservative with respect to Type-I errors, we report the most likely values for the test statistic and the maximal p-values across all runs.

The model revealed that among noun pairs, overall gender sameness was predicted by two composite factors: (1) the frequency of the words in the pairing; and (2) the semantic similarity of the pair modulated by their co-occurrence likelihood (Figure 3). In the case of (1), the model indicated that the lower the word frequencies of the pair, the more likely they were to belong to the same noun class ($\chi 2 = 599.38$; $p < 0.0001$). In the case of (2), it was found that the more tightly semantically coupled a pair of nouns, the more likely they were to share gender. However, this pattern was further modulated by the mutual information between the noun pair ($\chi 2 = 711.43$; $p < 0.0001$): When the likelihood that the two nouns systematically co-occurred together was low, the effect of semantic similarity was attenuated; conversely, as co-occurrence likelihood increased, the effect of semantic similarity grew stronger. Thus, a noun pair was most likely to share the same gender when its nouns were both highly informative of one another and also contextually very similar.

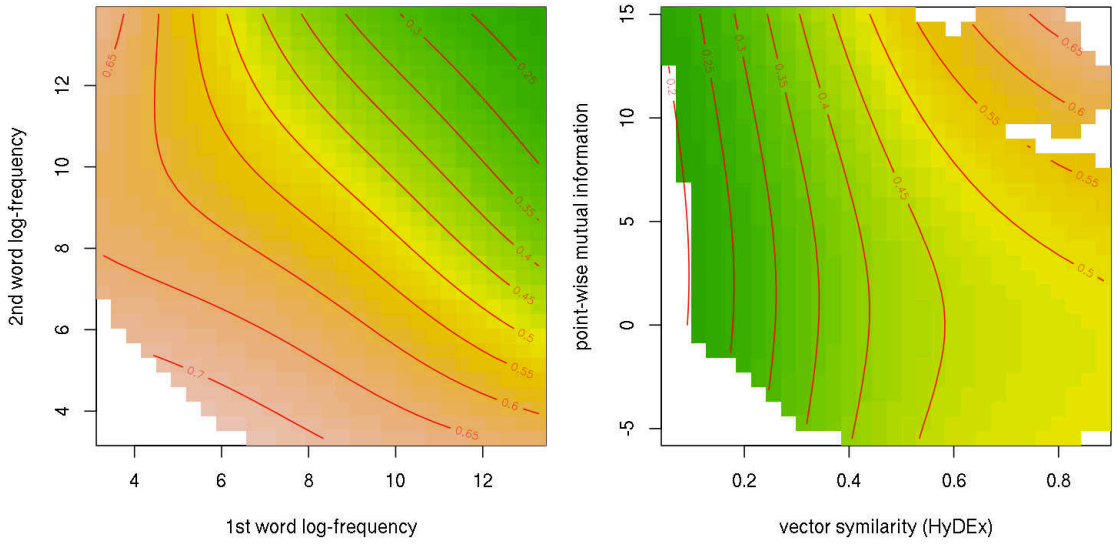

Fig 3. The final model revealed a complex pattern of effects with two numeric interactions (tensor

[4] To gather the necessary input to the model, the RFTagger was first run over the SdeWaC, expanding article contractions and lemmatizing noun forms (Schmid & Laws, 2008). Nouns that occur in the corpus in all case and number permutations were then selected for analysis (61K in total), with individual frequency tabulated as a lemma count, and co-occurrence rates between noun pairs calculated within a 2-word bidirectional window. These frequency and co-occurrence counts were used to compute PCDI, a measure of association that compares the probability of two nouns co-occurring against the probability of them occurring independently (Church & Hanks, 1989). Finally, the semantic similarity of noun pairs was calculated by running the High Dimensional Explorer (HiDEx: Shaoul & Westbury, 2010) over the lemmatized corpus, using a 5-word bidirectional window, and inverse linear ramp weighting. HiDEx is an implementation of the hyperspace analog to language (HAL) semantic space model, which stores raw lexical co-occurrence information in a high-dimensional matrix that it subjects to a series of transforms, yielding semantic similarity relations.

products): one between the noun's frequencies, and the second one between the mutual information and semantic similarity. Both tensor products were highly significant, and additional analyses of all possible partial effects reassured us that the terms in the model were strongly supported.

These results provide comprehensive quantitative support for the idea that there are systematic semantic trends in noun class assignment, indicating that while nouns that are semantically similar tend to belong to the same gender, this effect is modulated by frequency. Whereas high-frequency items tend to be distributed *across* genders, low-frequency items tend to be clustered within the *same* gender. Hence, the gender marking system in German appears to make use of both semantic clustering and semantic dispersion, with the choice of strategy varying with frequency.

An additional question worth pursuing is whether these strategies are realized differently in different classes. In fact, the probabilities of nominal gender in German differ markedly, with nearly half of nouns classed as feminine (49.45%), roughly a third as masculine (31.64%), and close to a fifth as neuter (18.96%). To assess whether nouns might pattern into different genders on the basis of their frequency, we attempted to predict noun class from noun frequency, using Bayesian multinomial logistic regression (*BayesLogit* package in R Statistical Computing Environment; see: Polson, Scott, & Windle, 2013; R Core Team, 2015).[5] This analysis revealed that noun frequency does *not* predict noun class (Figure 4). Thus, while there appear to be strong general biases in class assignment, these biases do not pattern by frequency, suggesting that the different classes likely share quite similar distributional properties.

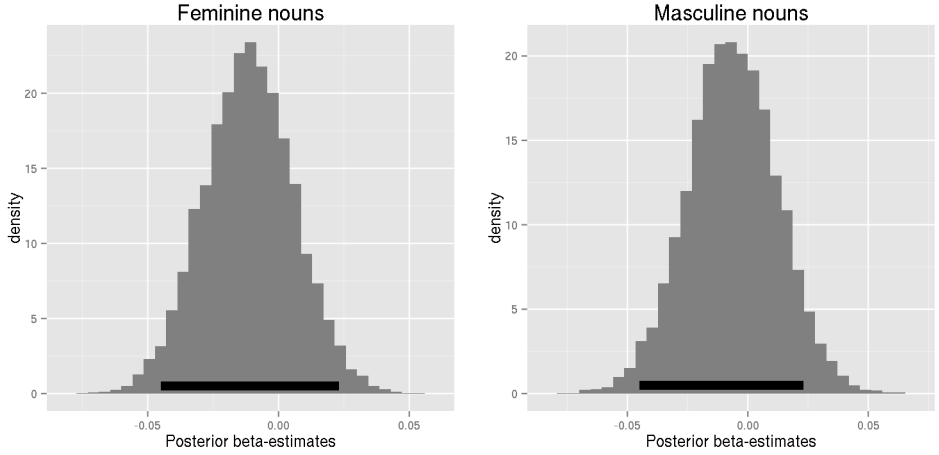

Fig 4. The posterior densities of the estimated coefficients for the frequency covariate for feminine (left panel) and masculine nouns (right panel). Bayesian credible intervals (95% HPD) are marked with a black horizontal line. As can be seen, the coefficients fall close to zero, and range over both positive and negative values. Such a result indicates that masculine and feminine nouns are distributionally indistinguishable from neuter nouns (the reference level in the model).

4 Why Taxonomy Misses the Point

[5] To make the computation feasible, the algorithm was run over a randomly selected sample of 12,000 nouns. Markov Chain Monte Carlo (MCMC) sampling was applied to obtain the posterior distribution of the regression parameters. The first 1,000 iterations were excluded as part of an initial burn-in, after which results were analyzed for 10,000 MCMC iterations.

The studies reported here provide evidence in support of our suggestion that German noun class is well-designed to help communicators predict nouns in context. The dispersal of nouns across different gender classes is clearly sensitive to factors that influence an item's discriminability, and appears structured to *level* the effects of these factors, making nouns more equally predictable in context.

While it has often been claimed that the German gender system is unsystematic and meaningless, our findings suggest, to the contrary, that not only does noun class serve to efficiently manage nominal entropy, but also that—like many other subsystems of language—gender in German is more specifically informative about high-frequency nouns than low-frequency nouns. It is notable that verb inflection, in both German and English, shares the same pattern: high-frequency verbs tend to have specific (irregular) inflection patterns that are highly informative about the inflected form of a given verb, whereas low-frequency verbs have generic (regular) inflection patterns that are less specifically informative (Baayen & Moscoso del Prado Martín, 2005).

While this point may seem counterintuitive, it is well predicted by a discriminative account. High and low-frequency forms pose markedly different challenges in terms of entropy management. Compared to lower frequency forms, high frequency items tend to be more contextually 'promiscuous' (Adelman, Brown, & Queseda, 2006), to be more semantically similar to more other words (Steyvers & Tenenbaum, 2005), and to have denser phonological neighborhoods (Andrews, 1992), meaning that they are, at once, less disambiguated by context, and more confusable with other items. At the same time, high-frequency items are more likely to be encountered in sparser, less distinctive linguistic contexts, where their prediction is unsupported by other material (Genzel & Charniak, 2002, 2003; Sigurd, Eeg-Olofsson & van de Weijer, 2004). Accordingly, since context will often fail to distinguish highly frequent (and thus highly likely) nouns, a great deal of uncertainty is inevitable, and a greater level of uncertainty reduction is called for.

The relation between gender and frequency also makes sense in terms of learnability: Rigid, highly informative conventions, such as gender marking, can only arise in a language if all of the speakers in a community reliably encounter and acquire them. In German, the distribution of nouns will support the learning of apparently 'arbitrary' gender markers for more common nouns, because by dint of their frequency in the input, these nouns will be encountered by young learners early in development, at a sensitive period in cortical maturation (cf. Thompson-Schill, Ramscar, & Chrysikou 2009). Accordingly, the learning of these forms will not be influenced by the top-down factors that inhibit the acquisition of irregulars in adults (Ramscar, Dye, & McCauley, 2013). By contrast, the rarity (or complete absence) of low-frequency nouns in child-directed speech will render the rote learning of their gender classes all but impossible (Blevins, Milin & Ramscar, *this volume*). Instead, the presence of converging semantic and acoustic cues will serve to make the gender of low-frequency nouns predictable (i.e., 'regular'; Frigo & McDonald, 1998). This neatly solves the problem of how to mark nouns in a system that, because of its highly skewed distribution, renders the task of learning new noun-forms and their classes a continuous process that is never complete—leaving the system supple and adaptable to the demands of learning across the lifespan (Ramscar, Hendrix, Love & Baayen, 2013; Ramscar, Hendrix, Shaoul, Milin & Baayen, 2014).

All of which is to say that the structure of German noun class is shaped by considerations that are the *opposite* of those that have traditionally been understood to determine gender assignment. Noun class serves a discriminatory purpose, and the information processing requirements of discriminatory and taxonomic systems are very

different. Masculine and feminine gender classes do not reflect any kind of deep underlying taxonomic distinction; rather, items are assigned to different gender classes because assigning them to different gender classes is systematically informative. While gender classification can appear to be taxonomically interpretable at a squint, up close, many of its classifications appear taxonomically senseless. Yet there is an underlying logic to the system that is evident throughout: Gender serves to redistribute the entropy of nouns, making them more predictable, on average, in context. When we at last dispense with the long-standing assumption that gender marking is taxonomic, we dispense too with the confusions that have plagued its study.

Acknowledgements

Many thanks are due to Christian Adam for his heroic feats of data collection in mining the various WaCky corpora (SdeWaC, ukWaC, hrWaC, etc).

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. 2006. Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science* 17(9). 814–823.

Andrews, S. 1992. Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18(2). 234–254.

Arnon, I. & Ramscar, M. 2012. Granularity and the acquisition of grammatical gender: How order of acquisition affects what gets learned. *Cognition* 122(3). 292-305.

Aylett, M., & Turk, A. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47(1). 31–56.

Baayen, R. H., Milin, P., Filipovic Durdevic, D., Hendrix, P., & Marelli, M. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118(3). 438-481.

Baayen, R.H. & Ramscar, M. 2015. Abstraction, storage, and naive discriminative learning. In Dabrowska, E. and D. Divjak (eds.), *Handbook of Cognitive Linguistics*, 99-120. De Gruyter Mouton.

Baroni, M. Bernardini, S., Ferraresi, A., & Zanchetta. E. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3). 209-226.

Bates, E., Devescovi, A., Hernandez, A., & Pizzamiglio, L. 1996. Gender priming in Italian. *Perception & Psychophysics* 58(7). 992–1004.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. 2008. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1). 92–111.

Blevins, J., Milin, P. & Ramscar, M. (*this volume*) Zipfian discrimination

Bloomfield, L. 1933. *Language*. New York: Holt.

Campbell, J. 1982. *Grammatical man: Information, entropy, language and life*. New York: Simon & Schuster.

Chen, Y. S., & Leimkuhler, F. F. 1986. A relationship between Lotka's law, Bradford's law, and Zipf's law. *Journal of the American Society for Information Science 37*(5). 307-314.

Church, K. W., & Hanks, P. 1989. Word association norms, mutual information, and lexicography. *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, USA: Association for Computational Linguistics. 76–83.

Clark, H. H., & Wasow, T. 1998. Repeating words in spontaneous speech. *Cognitive Psychology* 37. 201-242

Claudi, U 1985. *Zur Entstehung von Genussystemen: Überlegungen zu einigen theoretischen Aspekten, verbunden mit einer Fallstudie des Zande*. Hamburg: Buske.

Corbett, G. G. 1991. *Gender*. Melbourne: Cambridge University Press.

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. 2000. Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language* 42(4). 465–480.

Faaß, G., & Eckart, K. 2013. SdeWaC - A corpus of parsable sentences from the web. Gurevych, Iryna, Chris Biemann & Torsten Zesch (eds.): *GSCL 2013, LNCS 8105*. Heidelberg: Springer.

Frank, A. F., & Jaeger, T. F. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*.

Frigo, L., & McDonald, J. L. 1998. Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language* 39. 218–245.

Gahl, S., & Garnsey, S. 2006. Knowledge of grammar includes knowledge of syntactic probabilities. *Language* 82(2). 405–410.

Gahl, S., Yao, Y., & Johnson, K. 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66(4). 1–18.

Genzel, D., & Charniak, E. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics: Morristown, NJ. 199–206.

Genzel, D., & Charniak, E. 2003. Variation of entropy and parse tree of sentences as a function of the sentence number. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*: Sapporo, Japan. 65-72.

Goldman-Eisler, F. 1958. Speech production and the predictability of words in context. *The Quarterly Journal of Experimental Psychology* 10(2). 96–106.

Grosjean, F., Dommergues, J.Y., Cornu, E., Guillelmon, D., & Besson, C. 1994. The gender-marking effect in spoken word recognition. *Perception & Psychophysics* 56(5). 590–598.

Hahn, U., & Nakisa, R. C. 2000. German inflection: Single route or dual route? *Cognitive Psychology* 41(4). 313-360

Heath, J. 1975. Some functional relationships in grammar. *Language* 51. 89-104

Holzmann, G. J., & Pehrson, B. 1994. *The early history of data networks*. Wiley-IEEE Computer Society Press.

Jaeger, T. F. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1). 23–62.

Keller, F. 2009. The entropy rate principle as a predictor of processing effort: an evaluation against eye-tracking data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain. 317-324.

Kilarski, M. 2007. On grammatical gender as an arbitrary and redundant category. In Kilbee, D. (ed.), *History of Linguistics 2005: Selected papers from the 10th International Conference on the History of Language Sciences* (ICHOLS X). John Benjamins, Amsterdam. 24–36.

Köpcke, Klaus-Michael. 1982. *Untersuchungen zum Genussystem der deutschen Gegenwartssprache* (Linguistische Arbeiten 122). Tübingen: Niemeyer.

Koval', A. I. 1979. O značenii morfologičeskogo pokazatelja klassa v fula. In N. V. Oxotina (ed.) *Morfonologija i morfologija klassov slov v jazykax Afriki*. Moscow: Nauka. 5-100.

Kullback, S. & Leibler, R.A. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22 (1). 79–86.

Kuperman, V., & Bresnan, J. 2012. The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language* 66(4). 1–24.

Kunz, M. 1988. Lotka and Zipf: Paper dragons with fuzzy tails. *Scientometrics 13*(5-6). 289-297.

Lakoff, George. 1986. Classifiers as a reflection of mind. In Craig, C. (ed.), *Typological Studies in Language 7: Noun Classes and Categorization*, 13-51.

Levy, R. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126–1177.

Lew-Williams, C., & Fernald, A. 2007. Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, 18(3), 193–198.

MacWhinney, B., Bates, E., & Kliegl, R. 1984. Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior* 23(2). 127–150.

Mandelbrot, B. 1953. An informational theory of the statistical structure of language. *Communication Theory* 84. 486-502.

Manin, D. 2006. Experiments on predictability of word in context and information rate in natural language. *Journal of Information Processes* 6. 229-236

Maratsos, M. P. 1979. How to get from words to sentences. In D. Aaronson & R. Rieber (eds.), *Perspectives in Psycholinguistics*. Hillsdale, N.J.: Erlbaum

Miozzo, M., & Caramazza, A. 1999. The selection of determiners in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 25. 907–922.

Pellegrino, F., Coupé, C., & Marsico, E. 2011. A cross-language perspective on speech information rate. *Language* 87(3). 539–558.

Piantadosi, S.T., Tily, H., & Gibson, E. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9). 3526-9.

Pluymaekers, M., Ernestus, M. and Baayen, R. H. 2005. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62. 146-159.

Polson, N. G., Scott, J. G., & Windle, J. 2013. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association 108*(504). 1339-1349.

Priva, U. C. 2010. Constructing typing-time corpora: A new way to answer old questions. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.

Qian, T., & Jaeger, T. F. 2012. Cue effectiveness in communicatively efficient discourse production. *Cognitive Science* 36(7). 1312–1336.

Quine, W.V.O. 1951. Two dogmas of empiricism. *The Philosophical Review* 60. 20–43.

Ramscar, M. & Baayen, H. 2013. Production, comprehension and synthesis: A communicative perspective on language. *Frontiers in Language Sciences* 4. 233.

Ramscar, M., Dye, M. & McCauley, S. 2013. Error and expectation in language learning: The curious absence of 'mouses' in adult speech. *Language* 89(4). 760-793.

Ramscar, M., Hendrix, P., Love, B. & Baayen, H. 2013. Learning is not decline: The mental lexicon as a window into cognition across the lifespan*. The Mental Lexicon* 8(3). 450-481.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P. & Baayen, R.H. 2014. The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science* 6. 5-42.

Ramscar, M. & Port, R. 2015. Categorization (without categories). In E. Dawbroska & D. Divjak (Eds.), *Handbook of Cognitive Linguistics*. De Gruyter Mouton.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. 2010. The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science* 34. 909-957.

Reddy, M. J. 1979. The conduit metaphor: A case of frame conflict in our language about language. In *Metaphor and Thought*.

Rescorla, R.A., & Allan R. Wagner, A.R. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H., and Prokasy, W. F. (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Croft

Schmid, H. & Laws, F. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. *COLING 2008*, Manchester, Great Britain.

Schriefers, H. 1993. Syntactic processes in the production of noun phrases. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19. 841–850.

Schriefers, H., & Teruel, E. 2000. Grammatical gender in noun phrase production: the gender interference effect in German. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26(6). 1368–1377.

Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27. 379–423, 623–656.

Shaoul, C. & Westbury, C. 2010. Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods* 42(2). 393-413.

Sigurd, B, Eeg-Olofsson, M, van de Weijer J. 2004. Word length, sentence length and frequency–Zipf revisited. *Studia Linguistica* 58. 37–52.

Skut, W., B. Krenn, T. Brants., & H. Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* (ANLP), Morgan Kaufmann, San Francisco.

Sperber, D. & Wilson, D. 1996. *Relevance: Communication and cognition.* (2nd ed.) Wiley-Blackwell.

Steyvers, M., & Tenenbaum, J. B. 2005. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science* 29(1). 41–78.

Team, R. C. 2015. R: A language and environment for statistical computing. Vienna, Austria; 2014. *URL http://www. R-project. org*.

Temperley, D., & Gildea, D. 2015. Information density and syntactic repetition. *Cognitive Science*. doi: 10.1111/cogs.12215

Thompson-Schill, S., Ramscar, M., & Chrysikou, E. 2009. Cognition without control: when a little frontal lobe goes a long way. *Current Directions in Psychological Science* 8(5). 259-263.

Tily, H., & Piantadosi, S. 2009. Refer efficiently: Use less informative expressions for more predictable meanings. In *Proceedings of the Workshop on the Production of Referring Expressions: Bridging the Gap between Computational and Empirical Approaches to Reference*.

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., & Bresnan, J. 2009. Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition* 1(2). 147–165.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. 2005. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31(3). 443–467.

Van Son, R., & Pols, L. 2003. How efficient is speech? *Proceedings of the Institute of Phonetic Sciences* 25. 171-184.

Vigliocco, G., Antonini, T., & Garrett, M. F. 1997. Grammatical gender is on the tip of Italian tongues. *Psychological Science* 8(4). 314–317.

Vigliocco, G., Vinson, D.P., Paganelli, F., & Dworzynski, K. 2005. Grammatical gender effects on cognition: Implications for language learning and language use. *Journal of Experimental Psychology: General* 134. 501-520.

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. 2004. Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience* 16. 1272–1288.

Wittgenstein, L. 1953. *Philosophical investigations*. Oxford, England: Blackwell.

Wood, S.N. 2006. *Generalized additive models: an introduction with R*. CRC press.

Wood, S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(1). 3-36.

Zipf, G.K. 1935. *The psychobiology of language*. New York: Houghton-Mifflin.

Zipf, G.K. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley: Cambridge, MA.

Zubin, D & Köpcke, K-M. 1986. Gender and folk taxonomy: The indexical relation between grammatical and lexical categorization. In: C. Craig (ed.): *Noun Classification and Categorization* (Typological Studies in Language; Vol. 7). Philadelphia: Benjamins, North America. 139–180.

Zubin, D & Köpcke, K-M. 1996. Prinzipien für die Genuszuweisung im Deutschen.In: Lang, E. and G. Zifonun (eds.): *Deutsch typologisch*. Berlin: de Gruyter. 473–491.

Zubin, D & Köpcke, K-M. 2009. Gender control-lexical or conceptual? In: Steinkrüger and Krifka (eds.): *Trends in Linguistics: On Inflection*. Berlin: de Gruyter. 237-262.