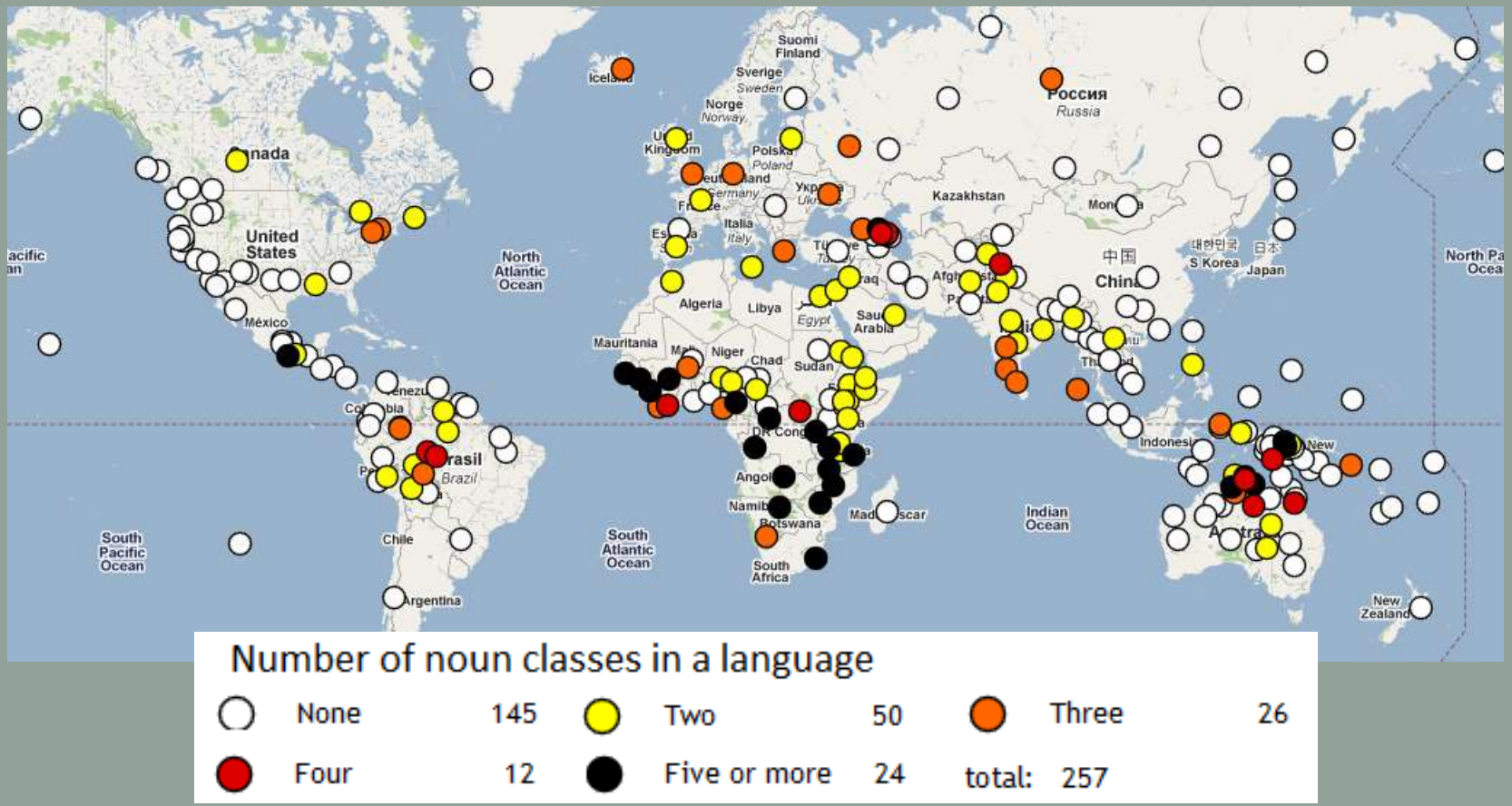
German Grammatical Gender Contributes to Communicative Efficiency

Richard Futrell and Michael Ramscar

January 2012

Noun class systems:
who needs 'em?



Corbett (2008)

What's it good for?

- Maratsos (1979: 235): "excellent testimony to the occasional **nonsensibleness** of the species"?
- Baudouin de Courtenay: "a **deformity** … responsible for … nightmares, pathological behaviour, erotic and religious delusions and sadism" (Kilarski 2007)?
- Lakoff (1986): Expressing **salient cultural categories**?
  - Or a **taxonomy** of categories?
- Zubin & Köpcke (1986: 173): reference tracking on **pronouns**?
  - Or expressing **coreference** i.e. on adjectives and verbs?
  -

Where we're going

- Grammatical gender **makes nouns more predictable** in context.
-
- 1. How does it work?
- 2. What are the effects?
- 3. Is it adapted for this function?

Where we're going

- Grammatical gender **makes nouns more predictable** in context.
-
- **1. How does it work?**
- 2. What are the effects?
- 3. Is it adapted for this function?

Reminder: Probability and information theory

- *P*(*e*) is the probability of event *e*.
  - For instance, the probability that a word will be "mansion" is the **frequency** of "mansion" divided by the **total frequency** of all words.
- *P*(*e* | *c*) is the probability of event *e* given context *c*.
  - For instance, the probability that a word will be "abode" given that it follows "humble".
  - P("abode") << P("mansion"),
  - But P("abode" | "humble") >> P("mansion" | "humble")!

Reminder: Probability and information theory

- We usually deal with the ***information*** of a probability, which is $-2\log_2 P(e)$.
  - **Low probability** = **high information** content
  - **High probability** = **low information** content
  - I.e., a predictable event is not informative, but an unpredictable event is very informative.
- The weighted average informativity of all possible events in a context is **entropy**, which measures **uncertainty** or **unpredictability**.

Reminder: information theory

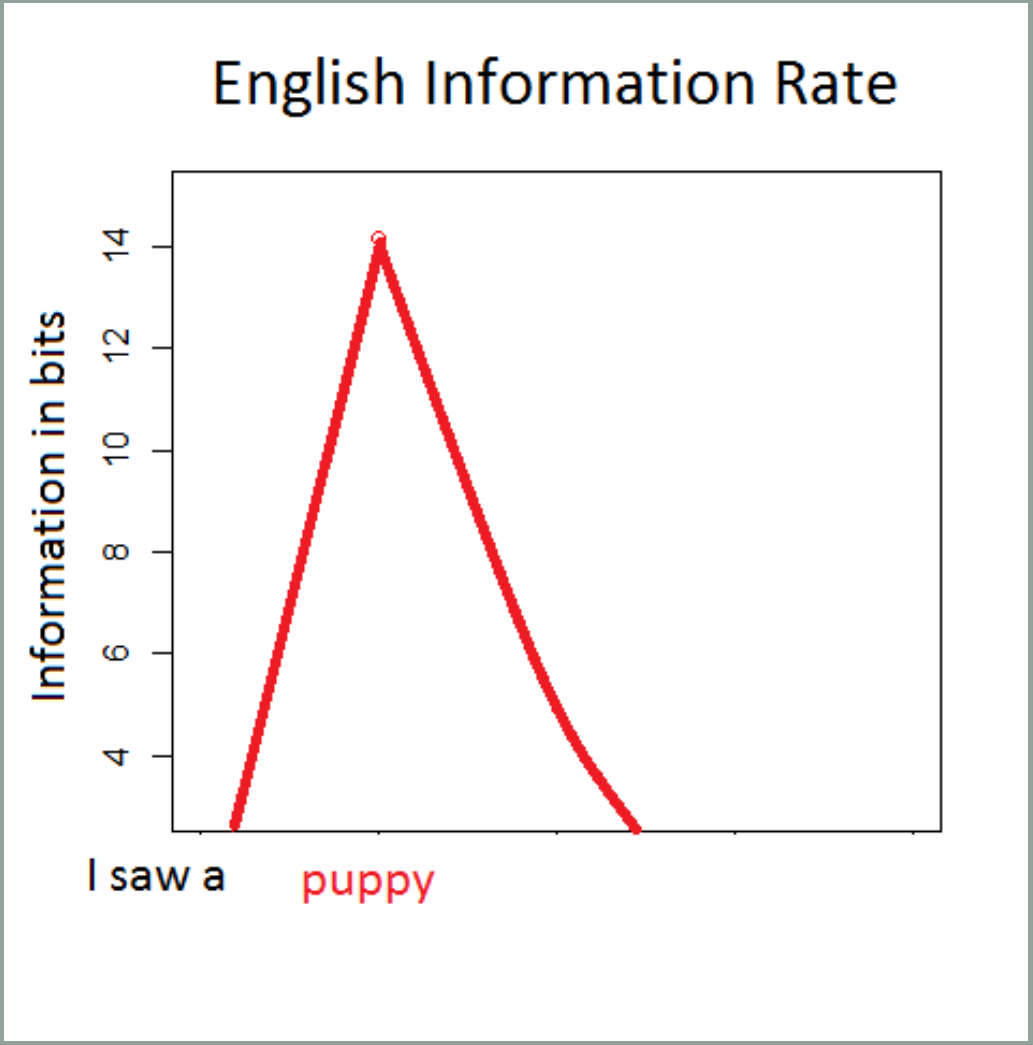
Wake up!

- Back to linguistics!

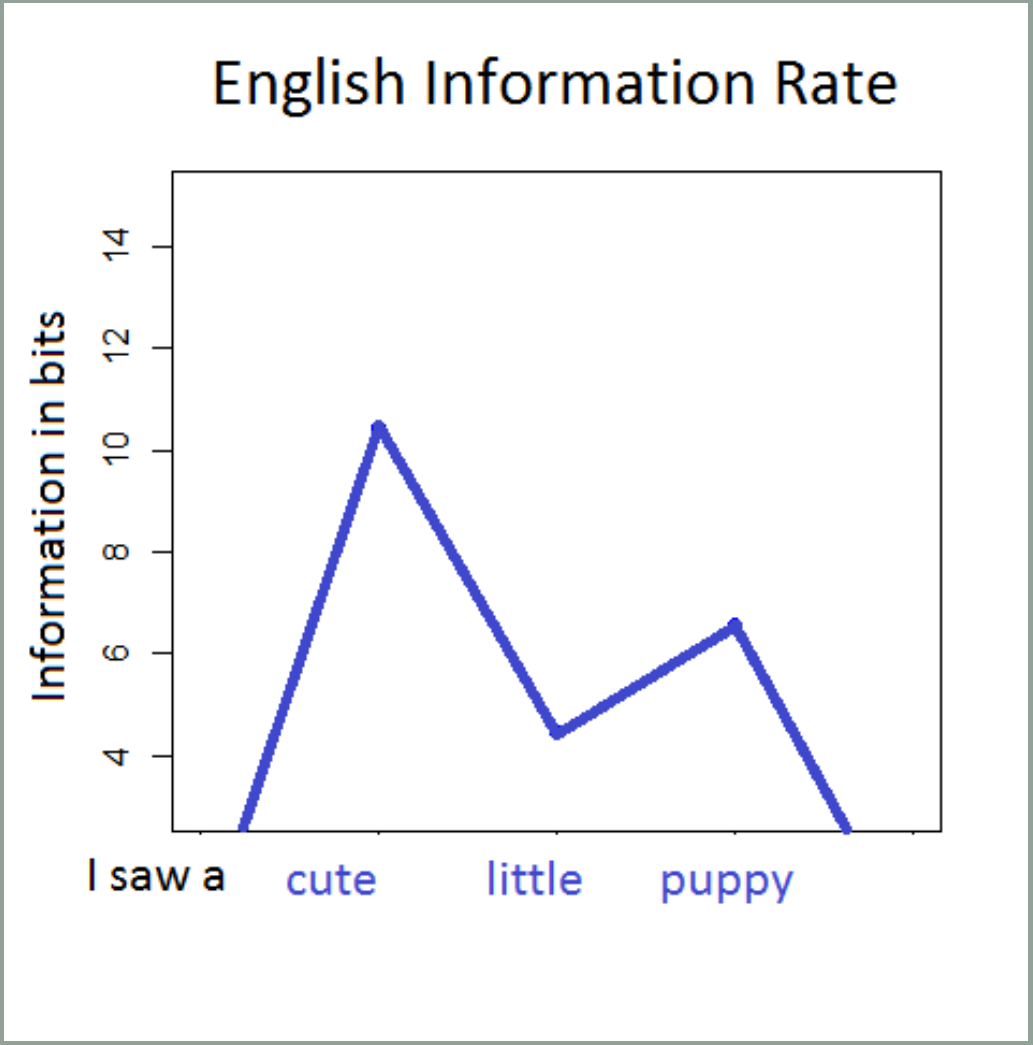
The trouble with nouns

- *Yesterday I saw a ! puppy.*
- *!* marks the site of disfluencies, pauses, and increased reading times, all linked with **uncertainty.**
- Unpredictable words represent **spikes in information flow**, which is inefficient (Jaeger 2008).
- Processing difficulties due to uncertainty are especially acute for nouns.
- The cure for **uncertainty** is **prediction**.

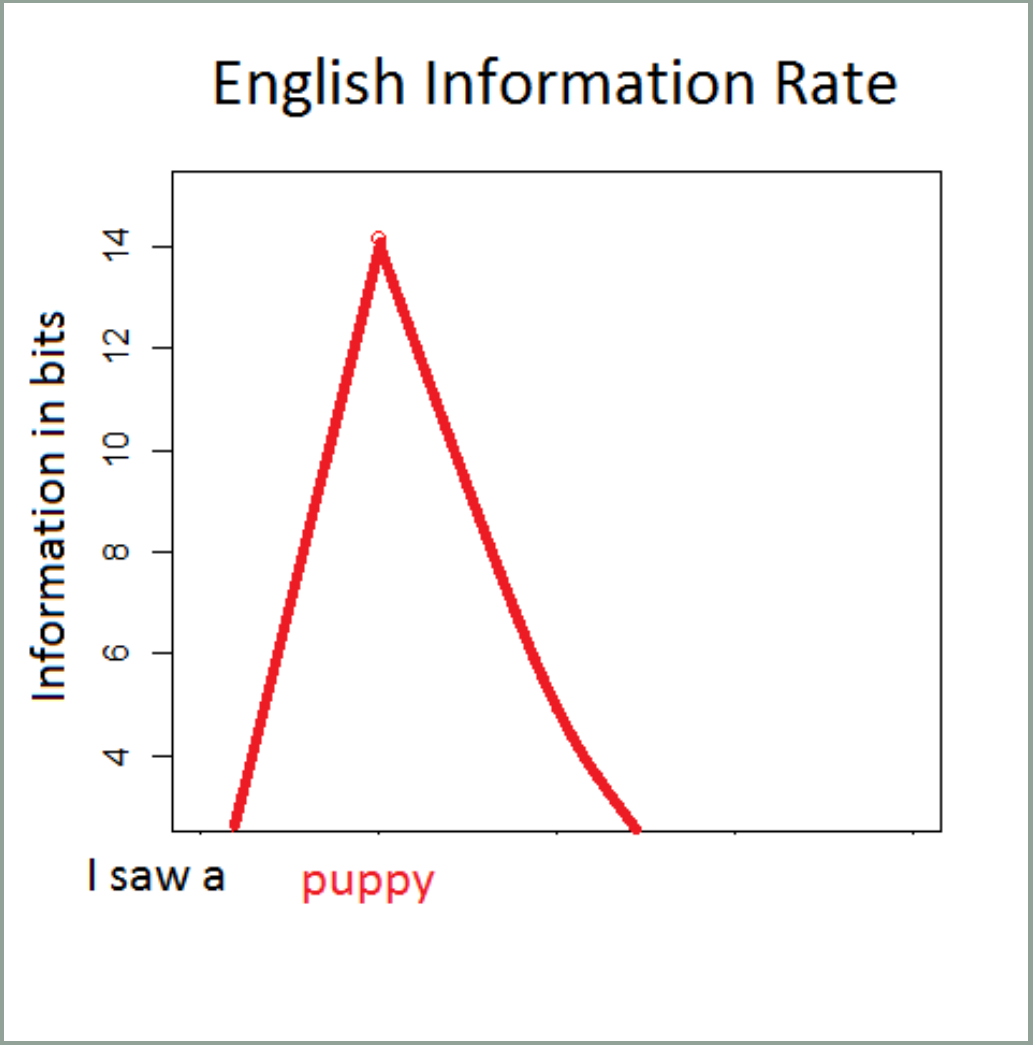
Fixing the trouble with nouns



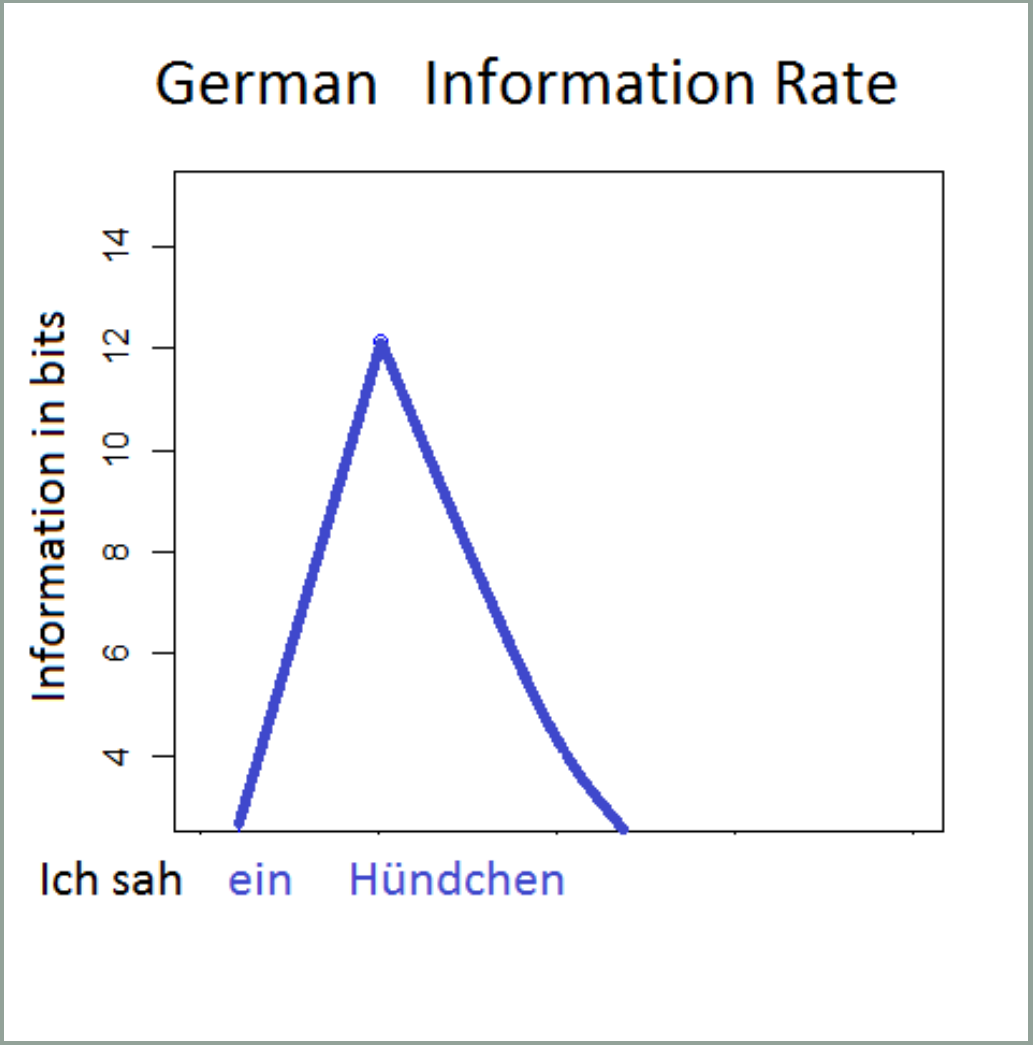
Fixing the trouble with nouns



Fixing the trouble with nouns



Fixing the trouble with nouns

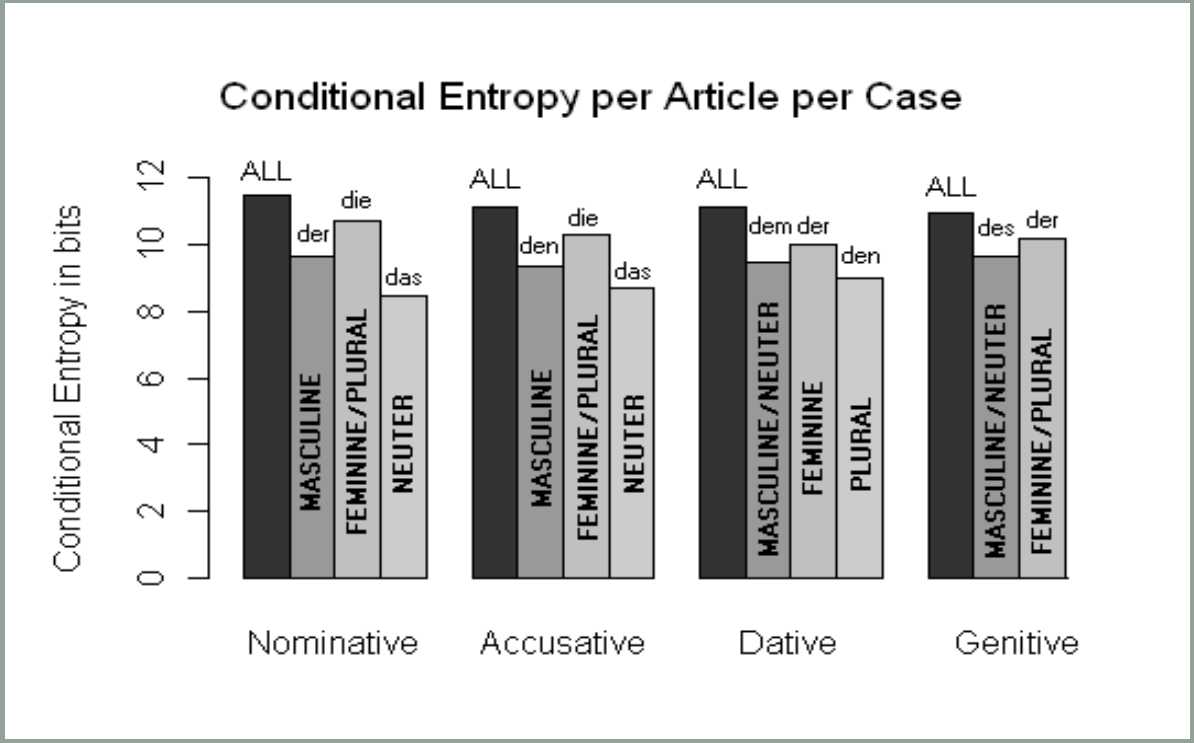


PREDICTIVE RELATIONS AND GENDER

- Noun class is an **extreme form** on a spectrum of possible probabilistic predictive relations.
- Genders are **almost totally predictable** given nouns.
- We can **describe** noun class as a predictive relation using **information theory** and the concept of **entropy.**
  - A language has a noun class system iff:
  - (1) There is some element *G* that co-occurs with nouns *N*,
  - (2) $H(N) > H(G) > 0$, and
  - (3) $H(G \mid N) \approx 0$.
  - This **implies that $H(N \mid G) < H(N)$**, i.e. that **gender predicts nouns**.

Predicting Nouns

- Noun class markers provide information about following nouns.
- Therefore they lower the entropy of those nouns: (based on counts from the NEGRA corpus of newspapers):



How gender works

- Gender is a form of **predictive context**. It can lower uncertainty about following words, or about pronominal reference, etc.
- The **information-theoretic description** captures gender as the extreme of a spectrum and it **makes gender's function obvious**.

Where we're going

- Grammatical gender **makes nouns more predictable** in context.
-
- 1. How does it work?
- **2. What are the effects?**
- 3. Is it adapted for this function?

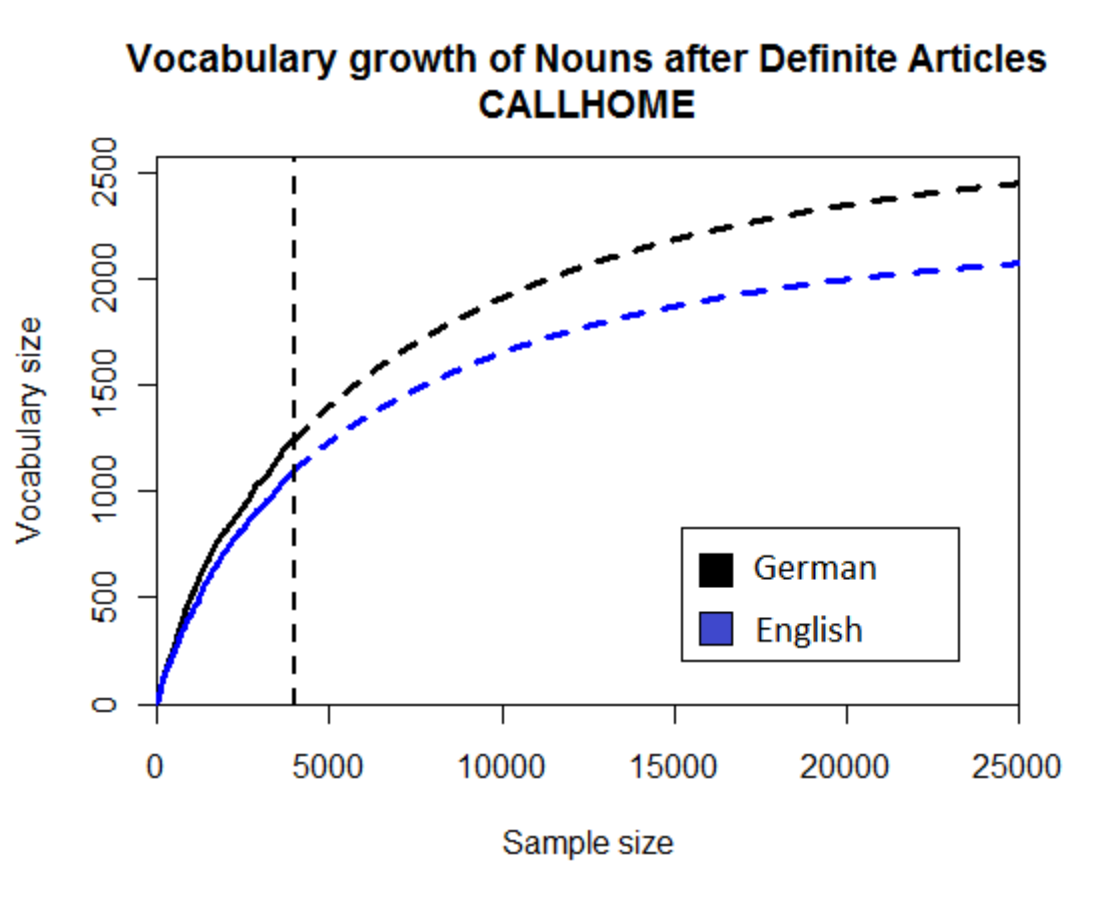
Gender makes infrequent nouns easier to use

- Germans should be **more comfortable** with **low-frequency** words like *Doberman* (as opposed to *dog*) directly after articles.
  - English speakers will make use of **other predictive context**, such as **adjectives**.
- So we should find **greater lexical richness** directly after articles in German than in English.

Measures of lexical richness

- **Vocabulary size** in a given sample (type-to-token ratio).
- **Vocabulary growth rate**, the rate at which the vocabulary grows as the sample size increases
  - **Baayen's *P***, the number of hapax legomena divided by sample size. Used as a measure of **morphological productivity**.
  - I measured these for nouns in the **CALLHOME Corpora** of spontaneous spoken German and English.

Spoken english and german



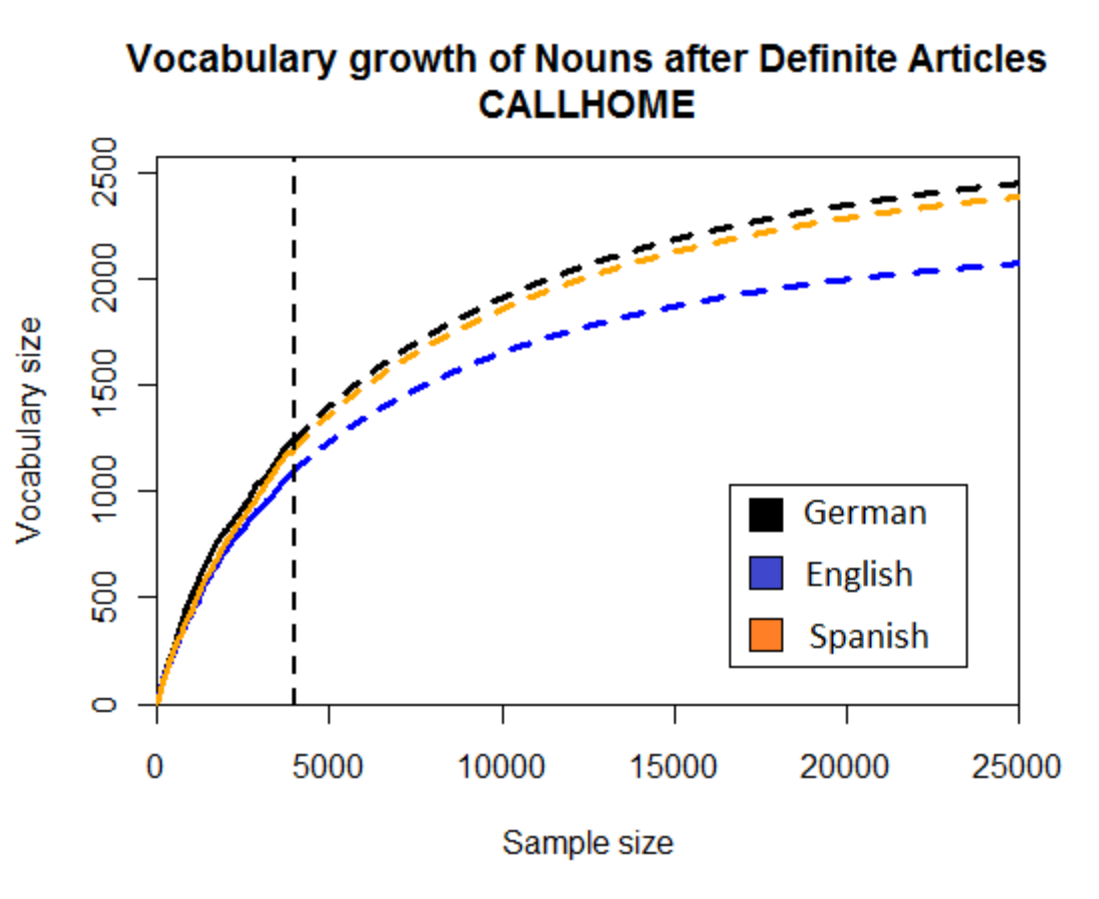
Spoken english and german

- two-tailed tests:
-

	z	p
Vocabulary Size	4.6142	0e+00
Vocabulary Growth Rate	3.3540	8e-04

- German H(N) = 11.7,
- German H(N|G) = 10.5,
- English H(N) = 10.1.

Spoken english, german, and spanish



Spoken english, german, and spanish

- two-tailed tests (Spanish vs. German)

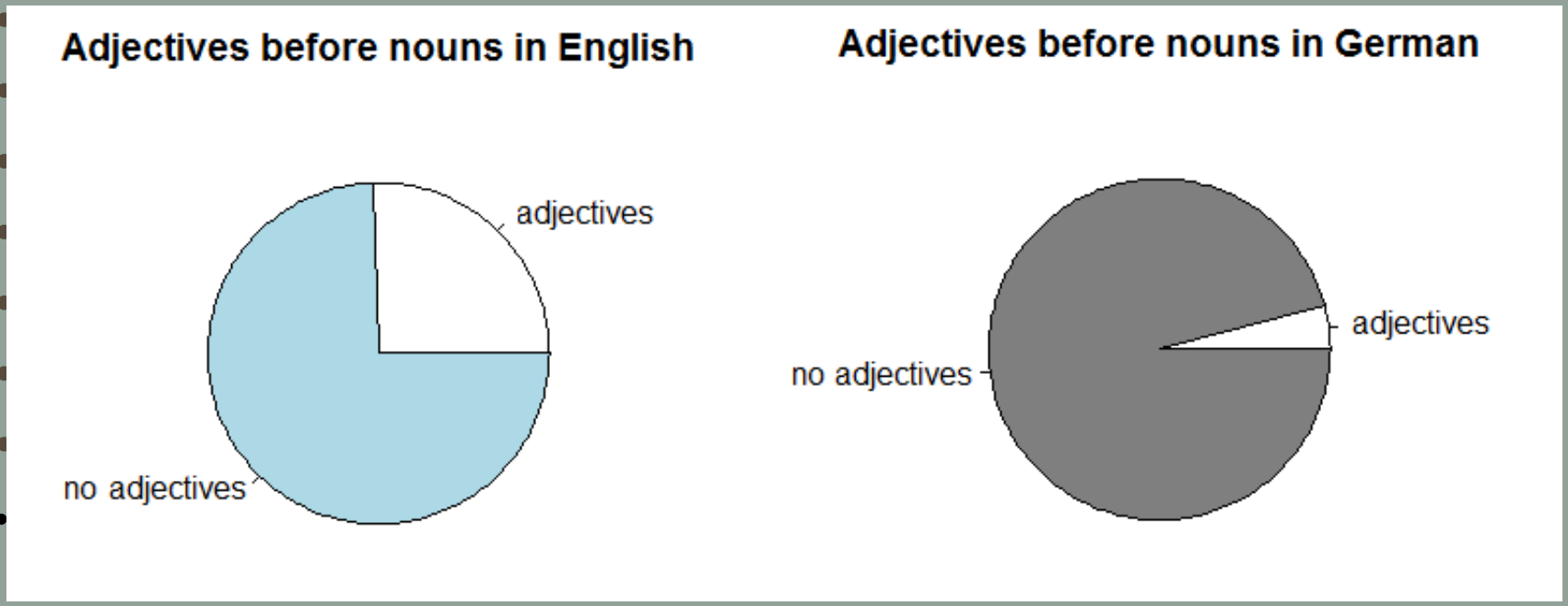
	Z	p
Vocabulary Size	1.0964	0.2729
Vocabulary Growth Rate	0.2752	0.7832

- two-tailed tests (English vs. Spanish)

	Z	p
Vocabulary Size	-3.4392	0.0006
Vocabulary Growth Rate	-3.0353	0.0024

How english makes up

•English speakers can still use infrequent nouns, but they need to be made predictable by **prenominal adjectives** (i.e. cute little puppy):



Lexical richness

- Gender marking allows for greater lexical richness of nouns.

Where we're going

- Grammatical gender **makes nouns more predictable** in context.
-

- **3. Is it adapted for this function?**

Gender assignment

- Gender can give cues about the semantic neighborhood of upcoming words.
- Zubin and Köpcke (1986) find regular semantic fields:
  - Weather conditions (masculine),
  - Alcoholic drinks (masculine),
  - Generic terms (neuter), etc.
- We aim to show semantic regularities quantitatively and comprehensively. **Semantically similar nouns should share gender**.

Measuring Semantic Similarity

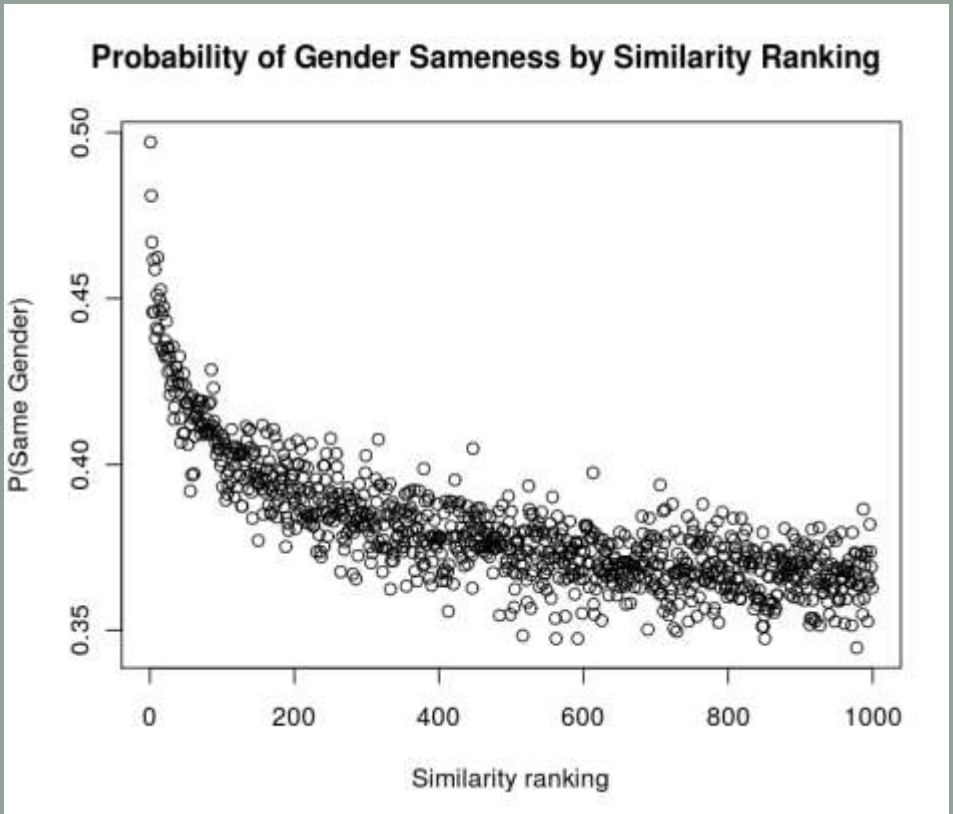
	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	0	0	1	1	1	0	1	0
information	0	0	1	1	1	0	1	0

vectors.

- Take the cosine between vectors as a measure of similarity
- Say that two words are similar if two vectors are similar
- I used this method in the **Google German 4-Gram corpus**, with gender information removed.

Predicting Noun Semantics

- Nouns are in the **same gender** as their **nearest semantic neighbor 50%** of the time:



Semantics and discrimination

- So, **semantically similar nouns** tend to be in the **same gender**. But not when **both are frequent.**

Where we went

- 1. How does it work?
  - Gender is predictable from nouns, so in context it **predicts nouns**.
- 2. What are the effects?
  - **Increased lexical richness** depending on availability of gender information.
- 3. Is it adapted for this function?
  - Yes—it facilitates the **prediction of the semantic neighborhood** of nouns.

Where we should go

- **More languages!** More corpora!
- Better psycholinguistic models, which should encompass the effects of **gender after nouns**.
  - Other concepts from communication/information theory: **error correction codes**?
- **Closer examination** of semantic/distributional effects. Does context predict (only) what gender doesn't?
- Extension of the theory to **classifier languages** like Chinese, and to other contextual predictive relationships.

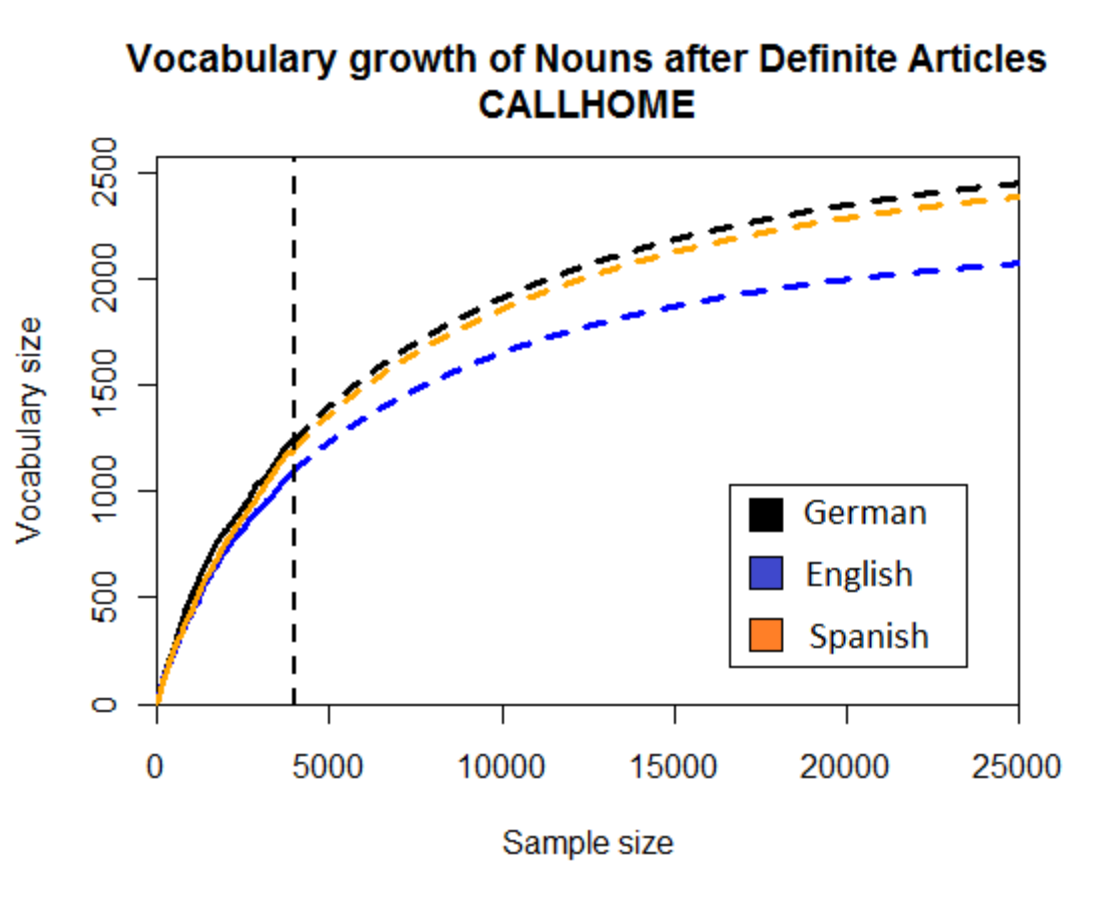
Acknowledgments

•We would like to thank **Dan Jurafsky** for patience and help, **Tom Wasow, Joan Bresnan,** and **the Spoken Syntax Lab** for thoughtful discussions, **Uriel Cohen Priva** for the example of conditional probability, **Victor Kuperman** for help with statistics and for constructive skepticism, **Hal Tily** and **Steven Piantadosi** for provocative suggestions, and Stanford Corpus TAs **Rob Munro**, **David Clausen**, and **Tyler Schnoebelen** for appeasing my neverending requests for corpora.

Works Cited

- **Baayen, R. Harald.** 2001. *Word frequency distributions*. Boston: Kluwer Academic Publishers.
- **Corbett, Greville G.** 2008. Number of Genders. In: Haspelmath, Martin & Dryer, Matthew S. & Gil, David & Comrie, Bernard (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 30. Available online at http://wals.info/feature/30 Accessed on 2010-05-20.
- **Kilarski M.** 2007. "On grammatical gender as an arbitrary and redundant category." In: Douglas Kibbee (ed.) History of linguistics 2005: Selected papers from the Tenth International Conference on the History of the Language Sciences (ICHOLS X), 1-5 September 2005, Urbana-Champaign, Illinois. Amsterdam - Philadelphia: John Benjamins, 24-36.
- **Lakoff, George.** 1986. Classifiers as a reflection of mind. In *Typological Studies in Language 7: Noun Classes and Categorization* (ed. Colette Craig), pp. 13-51.
- **Maratsos, M. P.** 1979. How to get from words to sentences. In D. Aaronson & R. Rieber (eds.), *Perspectives in psycholinguistics*. Hillsdale, N.J.: Erlbaum, 1979.
- **Zubin, D. A.** & **Klaus-Michael Köpcke.** 1986. Gender and folk taxonomy: the indexical relationship between grammatical and lexical categorization. In *Typological Studies in Language 7: Noun Classes and Categorization* (ed. Colette Craig), 139-80. Amsterdam: Benjamins.

Spoken english and german and spanish



Spoken english and german

- two-tailed tests (Spanish vs. German)

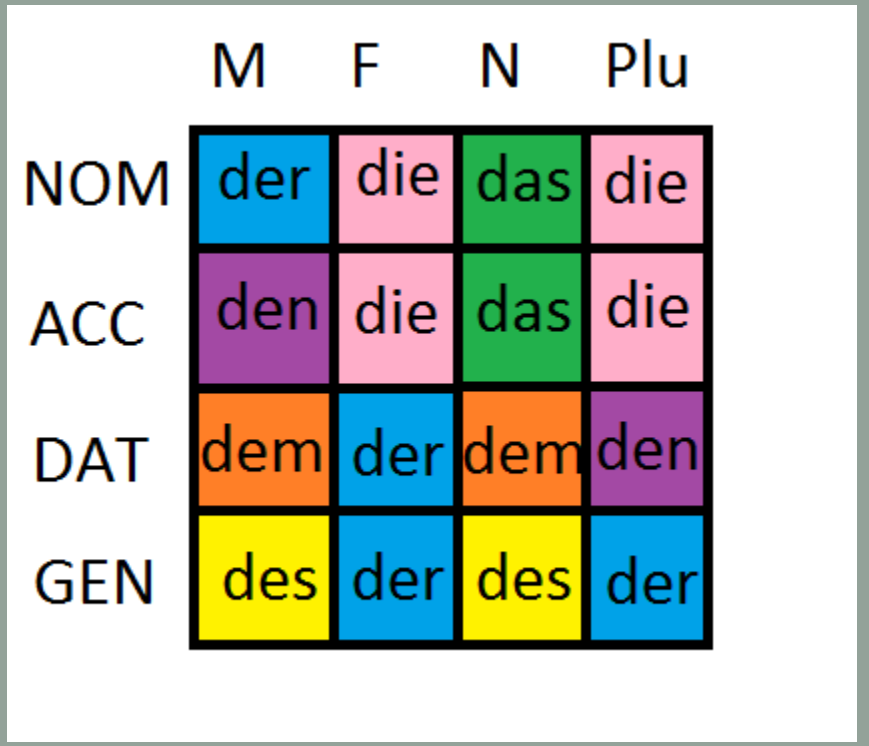
	Z	p
Vocabulary Size	1.0964	0.2729
Vocabulary Growth Rate	0.2752	0.7832

- two-tailed tests (English vs. Spanish)

	Z	p
Vocabulary Size	-3.4392	0.0006
Vocabulary Growth Rate	-3.0353	0.0024

Informativity of articles

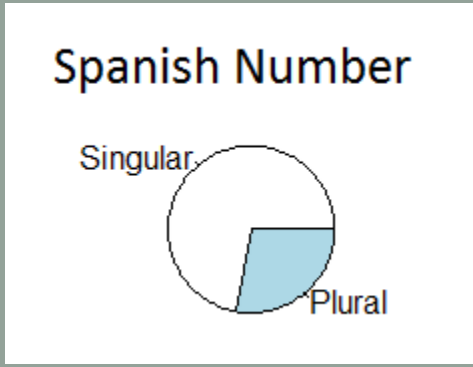
- Gender information is **always** available from Spanish articles.
- Gender information is **sometimes** available from German articles….
- Expected informativity of German articles: **1.04 bits**.
- Of Spanish articles: **1 bit.**

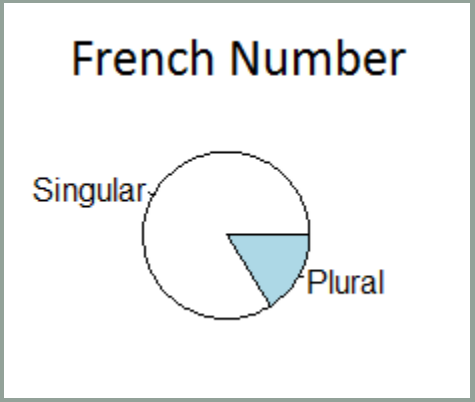


	M	F	N	Plu
NOM	der	die	das	die
ACC	den	die	das	die
DAT	dem	der	dem	den
GEN	des	der	des	der

Informativity of articles

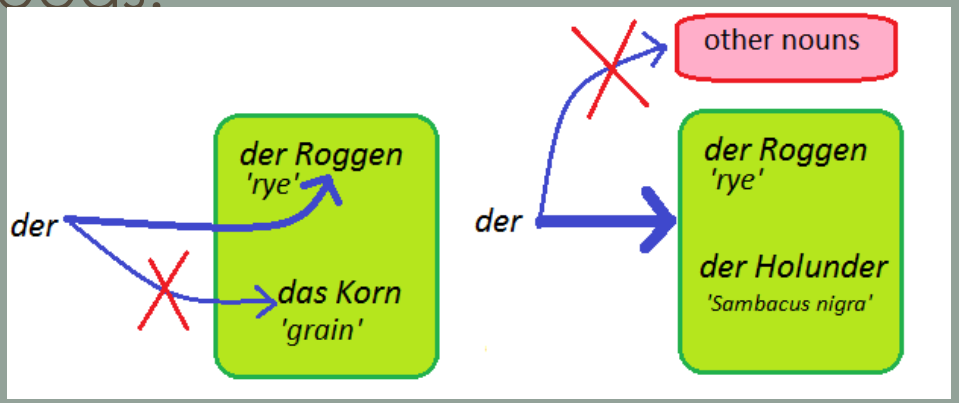
- French genders are **indistinguishable in the plural**.
- And French definite nouns are plural **16%** of the time.
- The expected informativity of French articles is **0.84 bits**.





two kinds of prediction

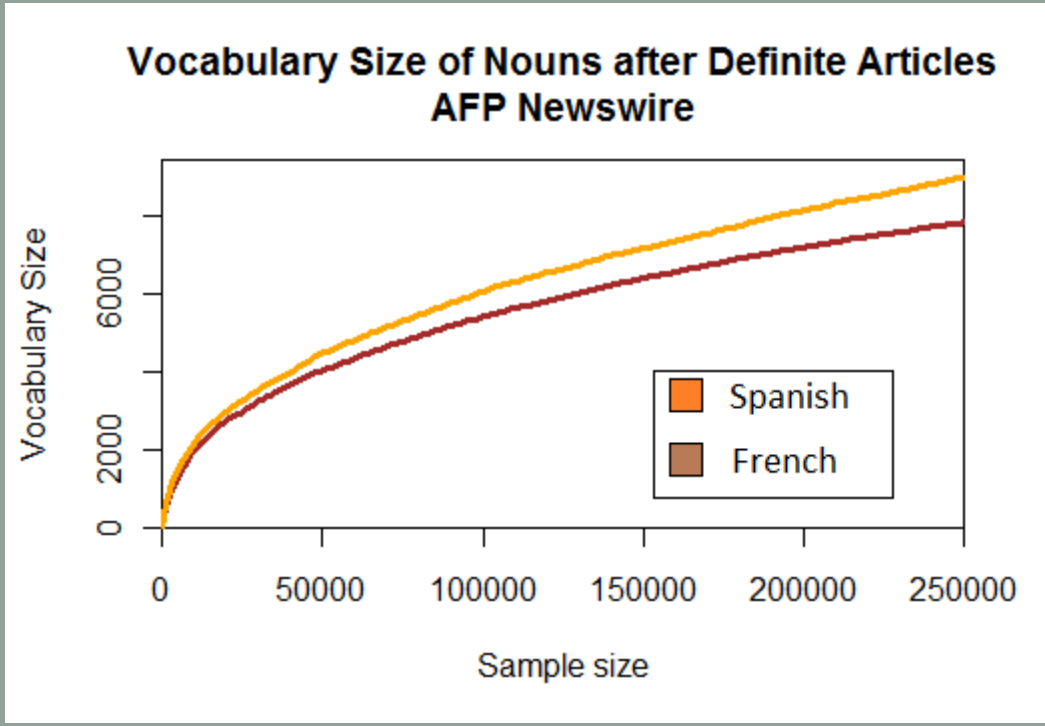
•Gender can either discriminate **within** members of a semantic neighborhood, or **between** semantic neighborhoods.



•If two nouns are (1) semantically **similar** and (2) **both frequent**, then speakers have to discriminate between them more frequently.

- •So they should be in different genders!

Spanish and French



two-tailed tests:

	Z	p
Vocabulary Size	-16.4796	0
Vocabulary Growth Rate	-10.1702	0

Informativity of articles