

EN ELLER ETT, GETTING THE SWEDISH GRAMMATICAL GENDER RIGHT

A probability based approach

GRAMMATICAL GENDER IN GERMAN

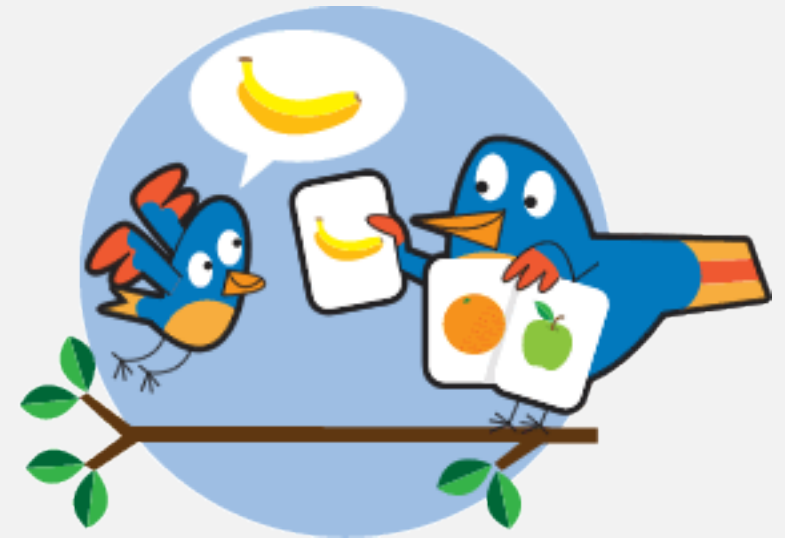
- Three gender system: Masc, Fem and Neuter
- Three articles: Der (m), Die (f), Das (n)
- The Table - **Der** Tisch (m)
- The Wall – **Die** Wand (f)
- The Car – **Das** Auto (n)
- Arbitrary?



HOW TO LEARN THE GERMAN GENDER?

- Intentional vs Incidental learning (Webb, 2019)
- Semantic and Morphologic Cues:
- **Der Mann – Die Frau – Das Kind**
- **Die Banane – Die Orange – Der Apfel**

But: **Der Käse – Das Ende**



A PROBABILISTIC APPROACH

- Tendencies, not rules!
- Distribution of genders:
m - 44%; f - 33%; n - 23% (Wagener, 1995)
- $\Pr(\text{DE} - m) = 0.44 \rightarrow$ Best guess always **Der**
- 90% ending -e = f (Köpcke, 1982) $\rightarrow \Pr(\text{DE} - f \mid -e) = 0.90$
- **Die** is a better guess!

___ Katze?



GENDER IN SWEDISH?

- Transparent language = many gender hints
- Two gender system: En (u) – Ett (n)
- **The** house – **das** Haus – **huset**
- **The** sun – **die** Sonne – **solen**
- (almost) arbitrary
- ~70% words **en (u)** (Bohnacker, 2003)
- $\text{Pr}(\text{SE} - u) = 0.70 \rightarrow 7:3$ odds!

Kupisch et al. 2022

<i>Spanish</i>	<i>Italian</i>	<i>Russian</i>	<i>French</i>	<i>German</i>	<i>Danish, Dutch, Norwegian, Swedish</i>
HIGH				Transparency	LOW

GETTING BETTER ODDS?

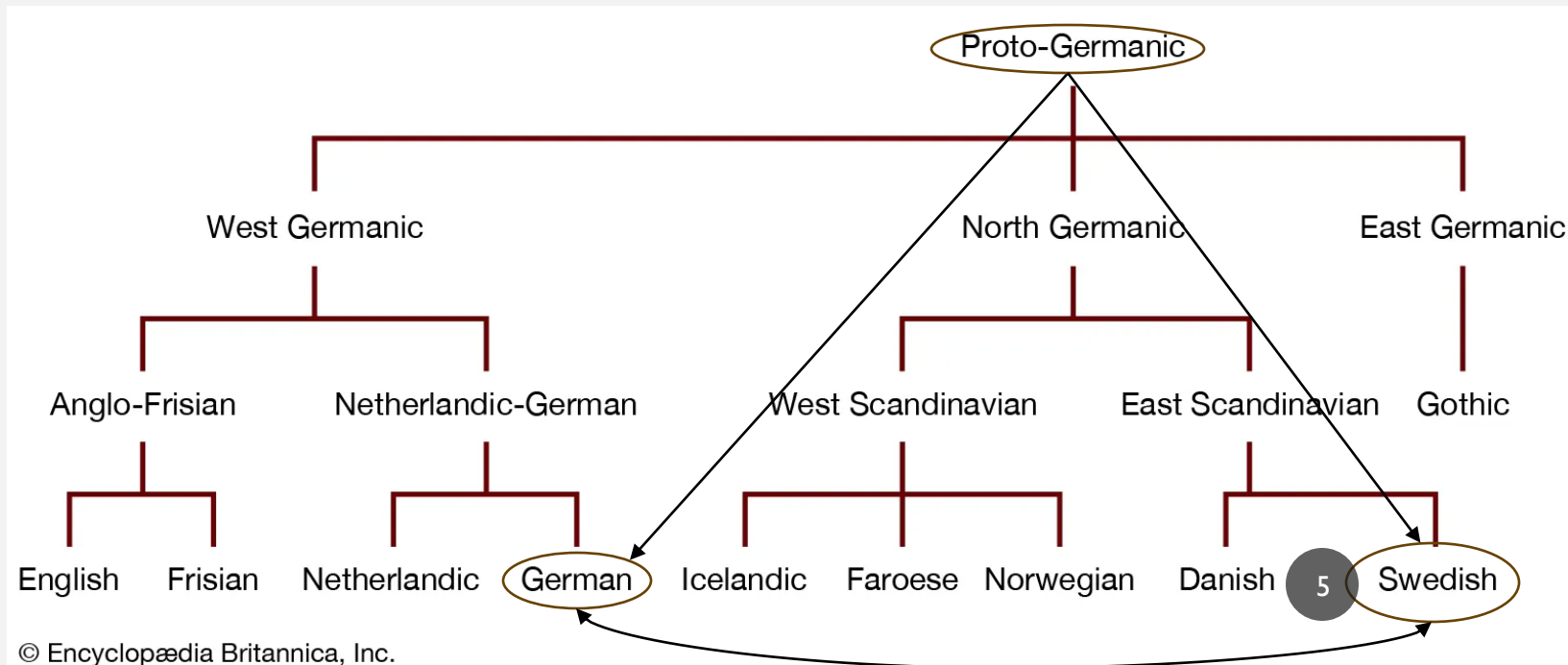
- Three gender system fell apart ~1500 - 1700 AD
- $m + f \rightarrow u$

Vad är Klockan? **Hon** är tolv

What is the Clock? **She*** is twelve

- **Die** Glocke/Uhr ($f = u$)

Relationship between German and Swedish gender assignment?



RESEARCH QUESTIONS

- Is there an association between gender assignment in Swedish and German?
- Is this influenced by shared genealogy (cognates)
- Can the relationship be used to improve the odds of guessing the grammatical gender of Swedish words?

Strategy 1: always guess utrum!

- $\Pr(\text{SE} - u) = 0.70$

Strategy 2: Always guess the German gender?

- $\Pr(\text{SE} - u \mid \text{DE} - u) = ?$
- $\Pr(\text{SE} - n \mid \text{DE} - n) = ?$

METHOD

- Sample from Parole Corpus (Språkbanken, 2017), N = 238
- Translated into German + encoded cognacy
- Join m + f → u
- Cross-tab analysis + Logistic Regression

SE_word	SE_article	SE_gender	DE_word	DE_article	DE_gender_u.n	cognate_status
gång	en	u	Gang	der	u	yes
dag	en	u	Tag	der	u	yes
köp	ett	n	Kauf	der	u	yes
man	en	u	Mann	der	u	yes
match	en	u	Match	das	n	yes
timme	en	u	Stunde	die	u	no
förändring	en	u	Veränderung	die	u	yes
år	ett	n	Jahr	das	n	yes
flytt	en	u	Umzug	der	u	no
lass	ett	n	Ladung	die	u	yes

Swedish	German	instances
n	n	1
n	u	2
u	n	1
u	u	6

RESULTS

Swedish ↓ German →	u	n	tot. swe
u	126	38	164
n	38	36	74
tot. ger	164	74	238

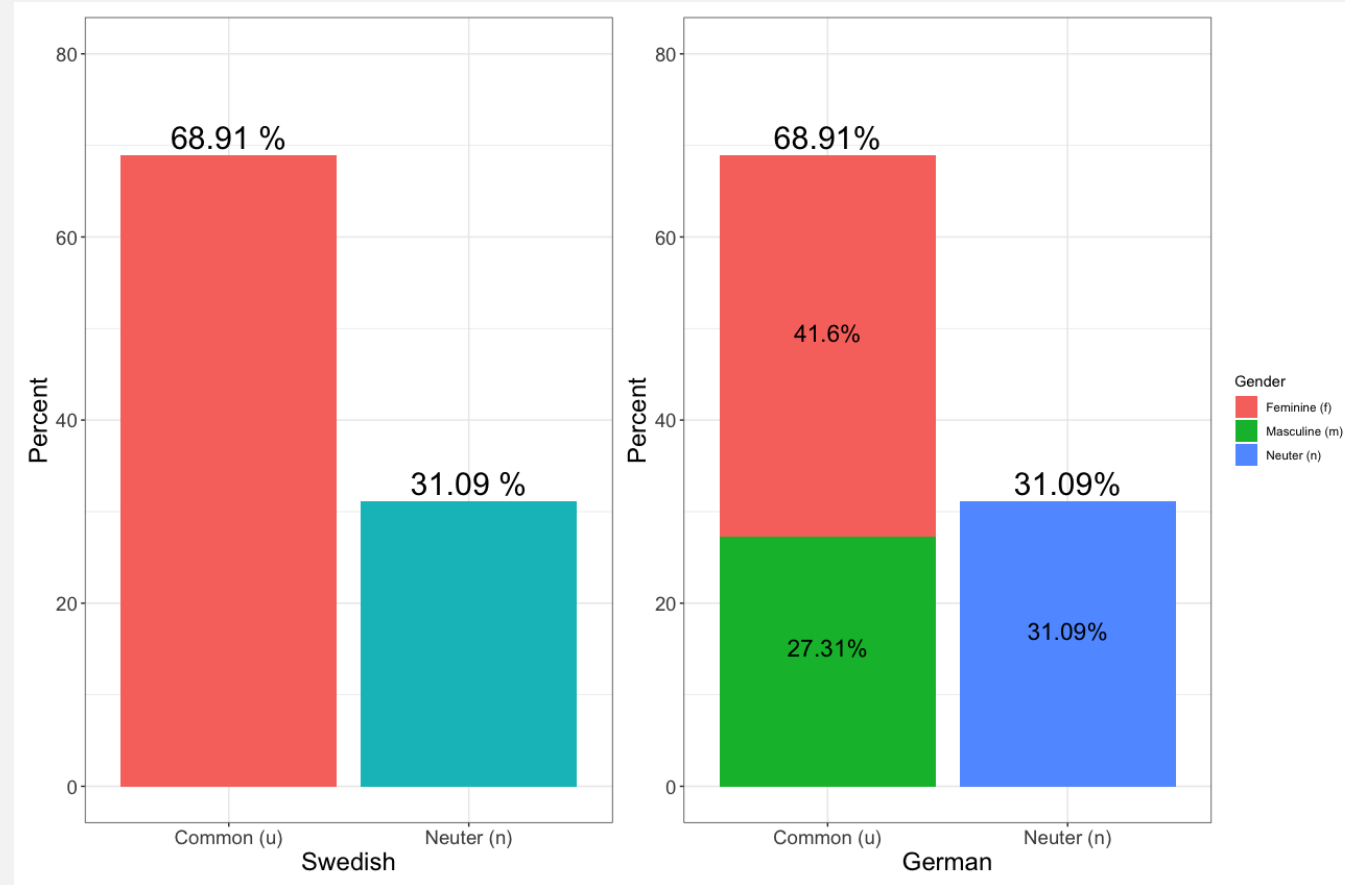
- $Pr(SE - u | DE - u) = \frac{126}{164} = 0.768$

- $Pr(SE - n | DE - n) = \frac{36}{74} = 0.486$

In this data:

- $E[strategy\ 1] = 0.689$

- $E[strategy\ 2] = 0.768 * 0.689 + 0.486 * 0.311 = 0.680$



Based on this data, ~ Strategy 1 gives marginally better results!

GENERAL CASE?

- Wagener(1995):
44% m, 33% f, 23% → 77% u, 23% n
- $Pr(SE - u | DE - u) = 0.768$
- $Pr(SE - n | DE - n) = 0.486$
- $E[Strategy 2] = 0.768 * 0.77 + 0.486 * 0.23 = 0.703$
- $E[Strategy 1] = 0.689$
- $E[Strategy 2] > E[Strategy 1]!$

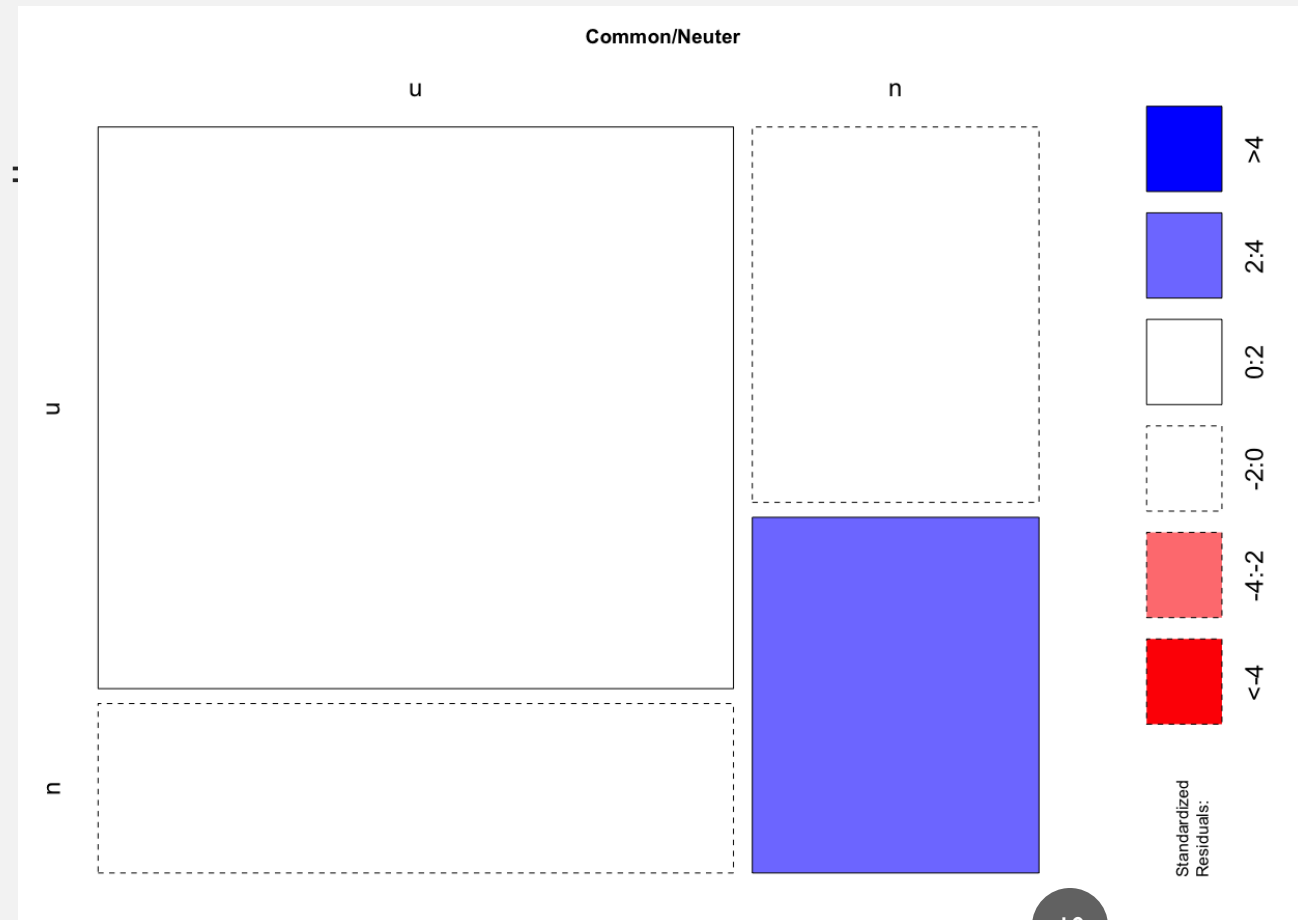
Conclusion: Depends on how you measure the baseline!



ASSOCIATION

- $SR = \frac{O - E}{\sqrt{E}}$
- $R_{nn} = O_{nn} - E_{nn} = 36 - \frac{74 \cdot 74}{238} = 36 - 23$
- $SR_{nn} = \frac{R_{nn}}{\sqrt{E_{nn}}} = \frac{13}{\sqrt{23}} = 2.71$
- Cramer's V = 0.255 → Medium effect

Swedish ↓ German →	u	n	tot. swe
u	126	38	164
n	38	36	74
tot. ger	164	74	238



MODELING

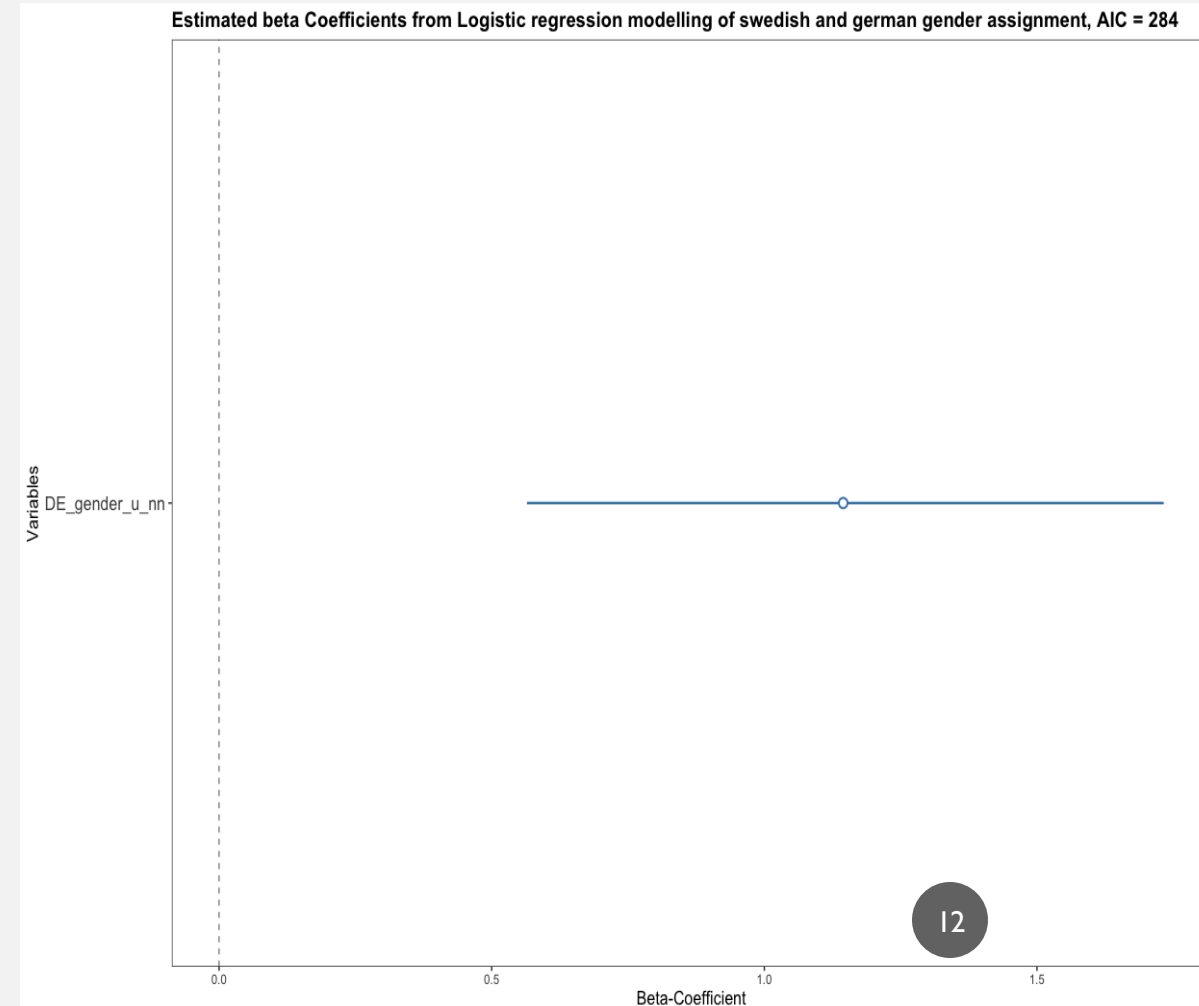
- Logistic Regression: $u = 0, n = 1$
- Target Variable: Swedish gender
- Predictor Variable: German gender
- OR = Odds ratio
- Control for cogancy!

$$\log\left(\frac{\Pr(SE = 1|DE)}{\Pr(SE = 0|DE)}\right) = \beta_0 + \beta_1 DE$$

$$OR = \frac{\text{odds}(x+1)}{\text{odds}(x)} = \frac{\left(\frac{p(x+1)}{1-p(x+1)}\right)}{\left(\frac{p(x)}{1-p(x)}\right)} = \frac{e^{\beta_0 + \beta_1(x+1)}}{e^{\beta_0 + \beta_1 x}} = e^{\beta_1}$$

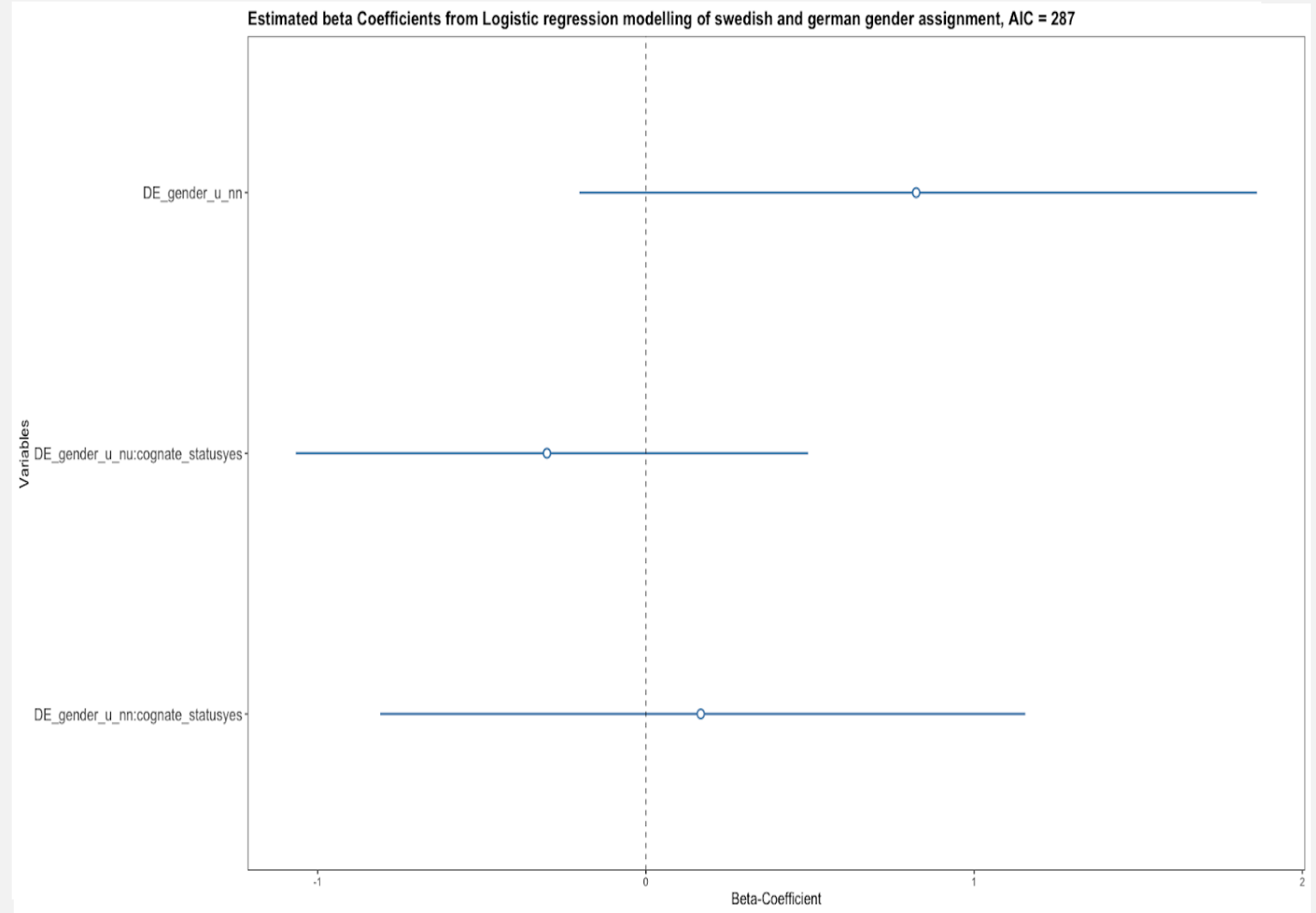
PARAMETER ESTIMATION

- Significant effect on Swedish gender assignment
 - $\beta_1 = 1.14$
 - $OR = e^{\beta_1} = e^{1.14} = 3.14$
 - 3 times higher odds if the gender is the same compared to if they were different!
- Guessing SE = u | DE = n is risky!



INFLUNCE OF COGNACY?

- No significant interacting effect
- Model worsened
- Gender assignment is not more likely to agree if the nouns are cognate



CONCLUSION

- Swedish and German gender assignment is moderately associated. Especially neuter words.
- Cognate words are not more likely to be assigned the same gender as non-cognates
- Large OR suggests taking German gender into account is important
- General strategy:
Always guess common gender, unless the German gender is neuter!

THANK YOU FOR LISTENING!

References:

Webb, S. (2019). *Incidental vocabulary learning*. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (1st ed., p. 15). Routledge. <https://doi.org/10.4324/9780429291586>

Köpcke, K. (1982). *Untersuchungen zum Genusystem der deutschen Gegenwartssprache*. Berlin, New York: Max Niemeyer Verlag. <https://doi-org.uaccess.univie.ac.at/10.1515/9783111676562>

Wegener, H. (1995). Das Genus im DaZ-Erwerb: Beobachtungen an Kindern aus Polen, Russland und in der Türkei. In B. Handwerker (Ed.), *Fremde Sprache Deutsch: Grammatische Beschreibung, Erwerbsverläufe, Lehrmethodik* (pp. 1–24). Tübingen: Narr.

Kupisch, T., Geiss, M., Mitrofanova, N. & Westergaard, M., (2022) “Structural and phonological cues for gender assignment in monolingual and bilingual children acquiring German. Experiments with real and nonce words”, *Glossa: a journal of general linguistics* 7(1). doi: <https://doi.org/10.16995/glossa.5696>

Språkbanken Text (2017). PAROLE (updated: 2017-05-17). [Data set]. Språkbanken Text. <https://doi.org/10.23695/x916-nm26>

Bohnacker, U. (2003). Nominal phrases. In: Josefsson, G.; Platzak, C. & Håkansson, G. (Eds.) *The Acquisition of Swedish Grammar*, 195–260. John Benjamins Pub

NO LOANS?

- Cramer's $V = 0.7$
- $OR = 50$
- Cognacy still no effect

