

Decoding Gender Assignment in a (Relatively) Opaque System: Experimental Evidence from Czech Nonwords

Mikuláš Preininger, Martina Abbondanza, Lukáš Jiříčka, Filip Smolík

May 2026, PsychoSlav, Zadar

Czech Academy of Sciences, Institute of Psychology

<https://earlygrammarlab.github.io>



Co-funded by
the European Union



Ministry of Education,
Youth and Sports
of the Czech Republic



Institute of Psychology
The Czech Academy of Sciences



<https://earlygrammarlab.github.io/people.html>



Martina Abbondanza



Lukáš Jiříčka



Filip Smolík

Arbitrary gender

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. (...) I translate this from a conversation in one of the best of the German Sunday-school books:

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.

Mark Twain: The Awful German Language

- Thousands of languages mark gender on nouns (Corbett, 1991)
- Biological sex is only the semantic core (she-turnip, he-moon)
- Languages differ in the transparency and complexity of the system

Systematicity & phonological cues

“[In German,] *monosyllabic words beginning with [kn], [dr] or [tr], with [s]+consonant, and those ending on a nasal+consonant, as well as those with two initial and two final consonants, **tend to be** masculine.*” (Szagun et al., 2007)

Systematicity & phonological cues

*“[In German,] monosyllabic words beginning with [kn], [dr] or [tr], with [s]+consonant, and those ending on a nasal+consonant, as well as those with two initial and two final consonants, **tend to be** masculine.” (Szagun et al., 2007)*

*“In the nominative singular, masculine nouns **typically** end in a consonant (e.g., dom ‘house’), **most** feminine nouns end in -a (e.g., lisá ‘fox’), and **most** neuter nouns end in -o (e.g., vedró ‘bucket’).” (Corbett, 1991)*

Systematicity & phonological cues

*“[In German,] monosyllabic words beginning with [kn], [dr] or [tr], with [s]+consonant, and those ending on a nasal+consonant, as well as those with two initial and two final consonants, **tend to be** masculine.”*
(Szagun et al., 2007)

*“In the nominative singular, masculine nouns typically end in a consonant (e.g., dom ‘house’), **most** feminine nouns end in -a (e.g., lisá ‘fox’), and **most** neuter nouns end in -o (e.g., vedró ‘bucket’).”* (Corbett, 1991)

*“[In Spanish,] words that end in –a, -d, ción, -sis, and –itis will be feminine in **98% of the time**, whereas words ending in in **any other grapheme** (-o, -l, -r, -n, -s, etc.) will be assigned a masculine gender **96% of the time**”* (Bull, 1965)

Systematicity & phonological cues

“[In German,] *monosyllabic words beginning with [kn], [dr] or [tr], with [s]+consonant, and those ending on a nasal+consonant, as well as those with two initial and two final consonants, **tend to be** masculine.*”
(Szagun et al., 2007)

“*In the nominative singular, masculine nouns typically end in a consonant (e.g., dom ‘house’), **most** feminine nouns end in -a (e.g., lisá ‘fox’), and **most** neuter nouns end in -o (e.g., vedró ‘bucket’).*” (Corbett, 1991)

“[In Spanish,] *words that end in -a, -d, ción, -sis, and -itis will be feminine in **98% of the time**, whereas words ending in in **any other grapheme** (-o, -l, -r, -n, -s, etc.) will be assigned a masculine gender **96% of the time***” (Bull, 1965)

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

Gender and morphology

- affixes can drive gender assignment
 - *-ción* is masculine while *-sis* is feminine in Spanish
 - *-ismus* is masculine while *-ung* is feminine in German

GG system of Czech nouns

FEMININE

žena

růže

píseň

kost

MASC ANIM

pán

muž

předseda

soudce

MASC INAN

hrad

stroj

NEUTER

město

moře

kuře

stavení

Questions

How do speakers deal with the degrees of systematicity?

What formal cues do they use when assigning novel words into gender classes and how do they use them?

- Experiment 1: explicit task
- Experiment 2, 3 & 4: implicit task

Exp 1: materials

	-ch	-n	-r	-t
non-palatalized consonant	<i>tech</i>	<i>lon</i>	<i>dyr</i>	<i>čut</i>
palatalized consonant	<i>teš</i>	<i>loň</i>	<i>dyř</i>	<i>čuť</i>
consonant + -a	<i>techa</i>	<i>lona</i>	<i>dyra</i>	<i>čuta</i>
consonant + -ka	<i>teška</i>	<i>lonka</i>	<i>dyrka</i>	<i>čutka</i>
consonant + -ek	<i>tešek</i>	<i>lonek</i>	<i>dyrek</i>	<i>čutek</i>

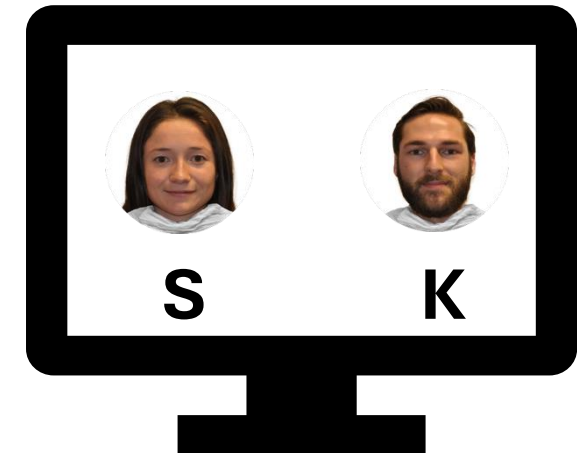
N = 100 (*dyr, loň, čuta, kmach, speř, cvina, vlotek, bruška, lopara, lasenek, tamitka, puloch, locuř, kobanka, kypet, joliš, radora, zalunek, sať, techa, dyrek, lonka, čut, kmašek, ...*)

Exp 1: procedure & participants

	-ch	-n	-r	-t
non-palatalized consonant	<i>tech</i>	<i>lon</i>	<i>dyr</i>	<i>čut</i>
palatalized consonant	<i>teš</i>	<i>loň</i>	<i>dyř</i>	<i>čuť</i>
consonant + -a	<i>techa</i>	<i>lona</i>	<i>dyra</i>	<i>čuta</i>
consonant + -ka	<i>teška</i>	<i>lonka</i>	<i>dyrka</i>	<i>čutka</i>
consonant + -ek	<i>tešek</i>	<i>lonek</i>	<i>dyrek</i>	<i>čutek</i>

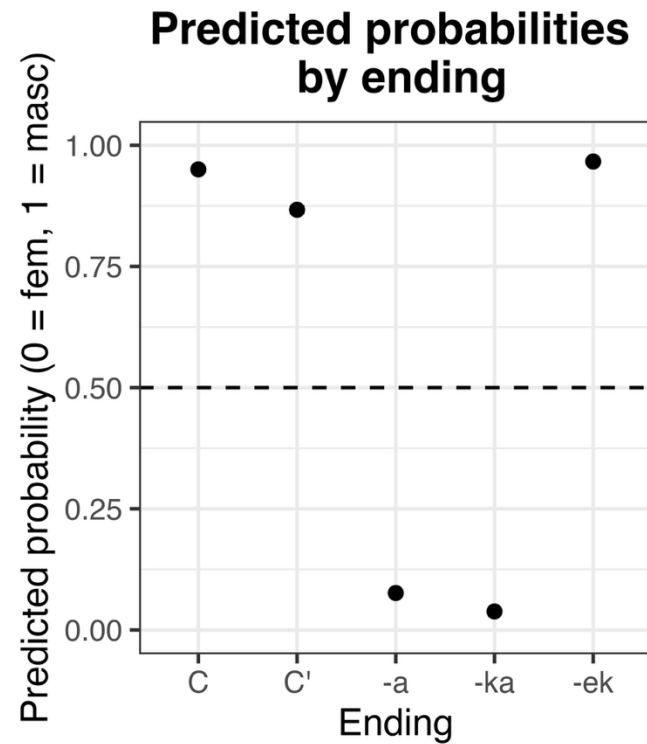
N = 100 (*dyr, loň, čuta, kmach, speř, cvina, vlotek, bruška, lopara, lasenek, tamitka, puloch, locuř, kobanka, kypet, joliš, radora, zalunek, sať, techa, dyrek, lonka, čut, kmašek, ...*)

 **jolich**



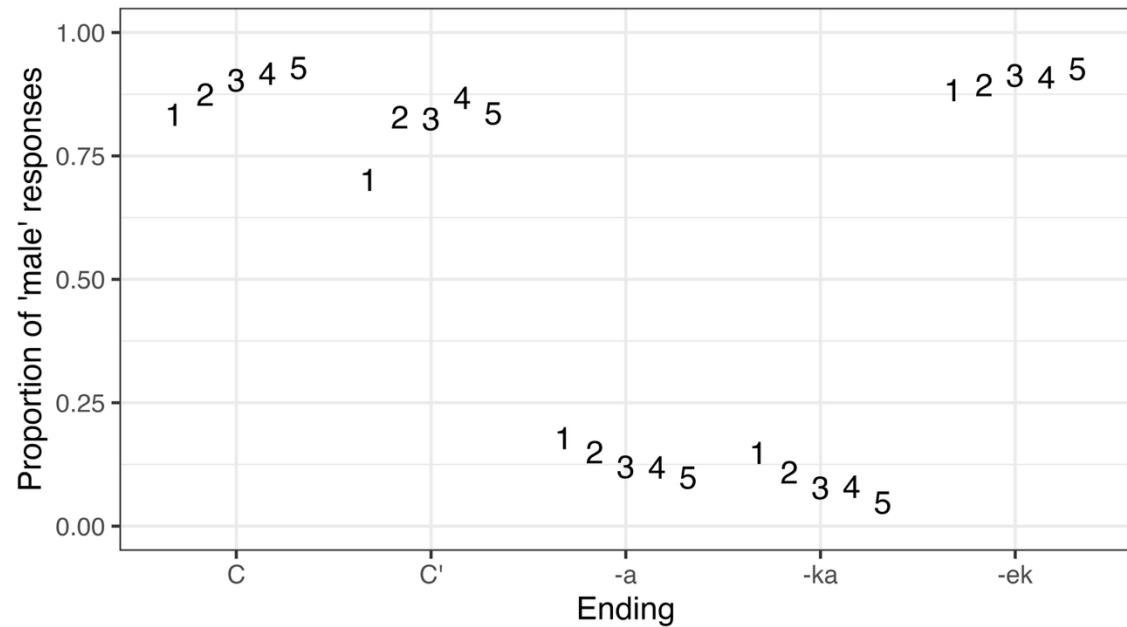
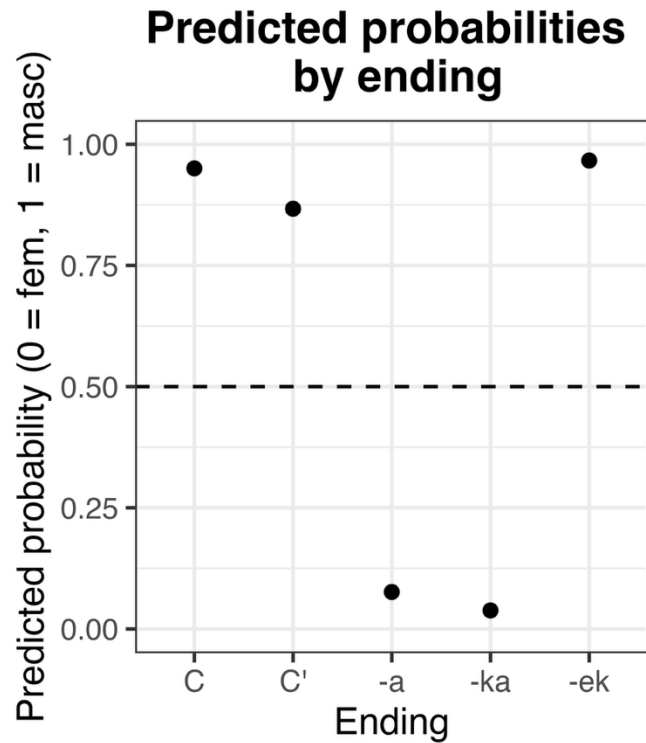
- online experiment
- “future roles/professions”
- N = 128 (107f, 20m, 1u)
- $M_{\text{age}} = 22.8$, $SD_{\text{age}} = 4.4$

Exp 1: results



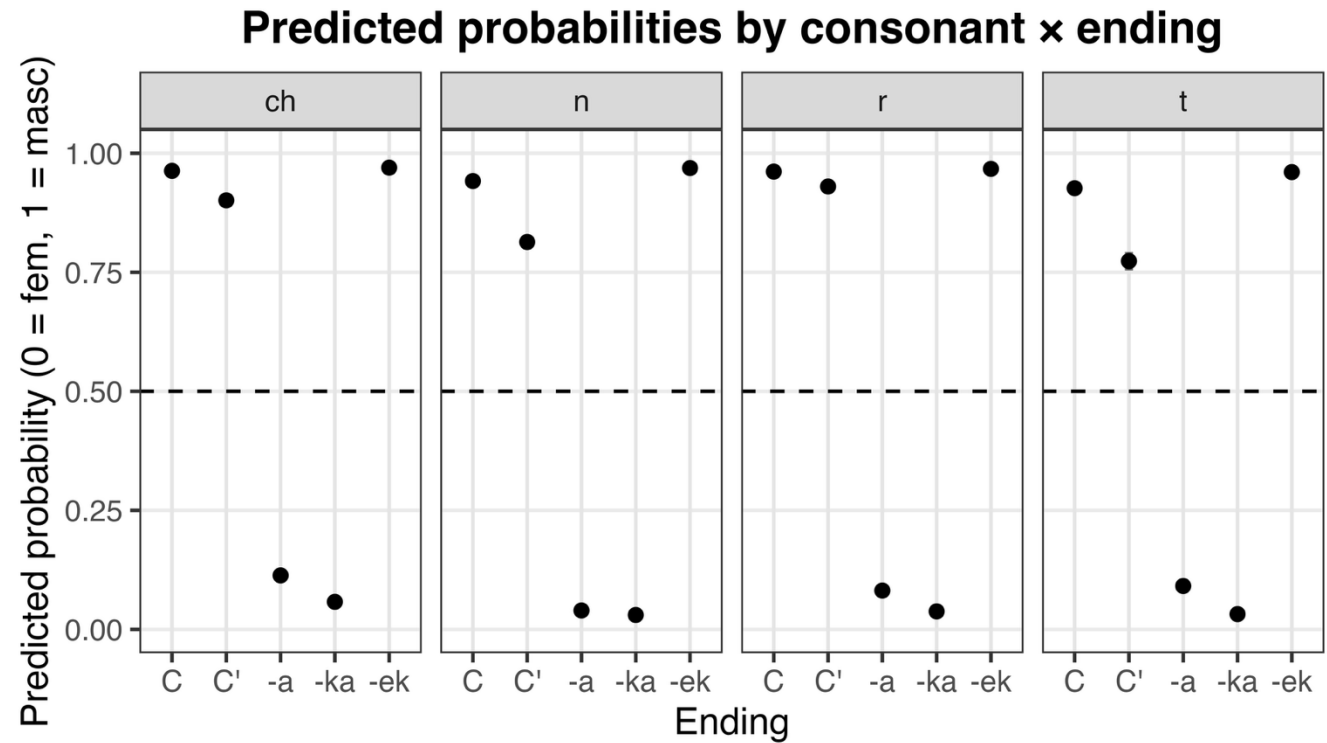
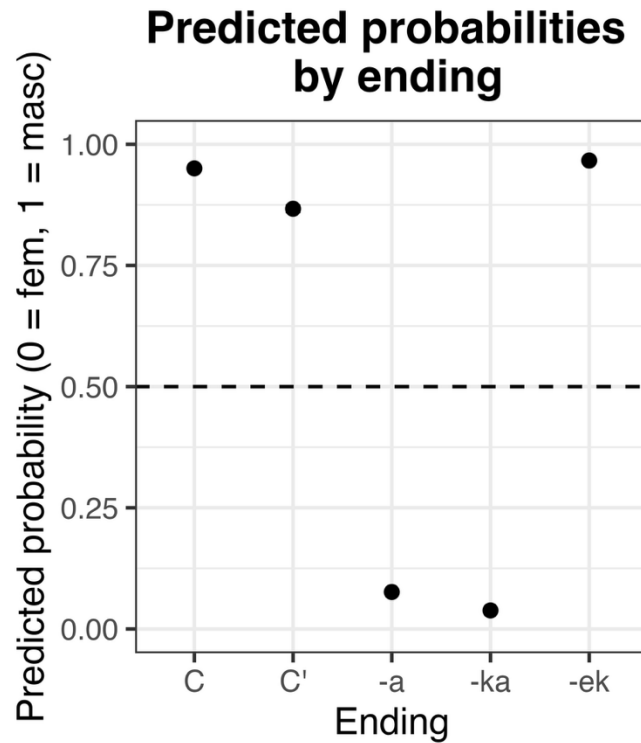
```
glmer(response_numeric ~ ending + (1 | part_id) + (1 + ending | base))
```

Exp 1: results



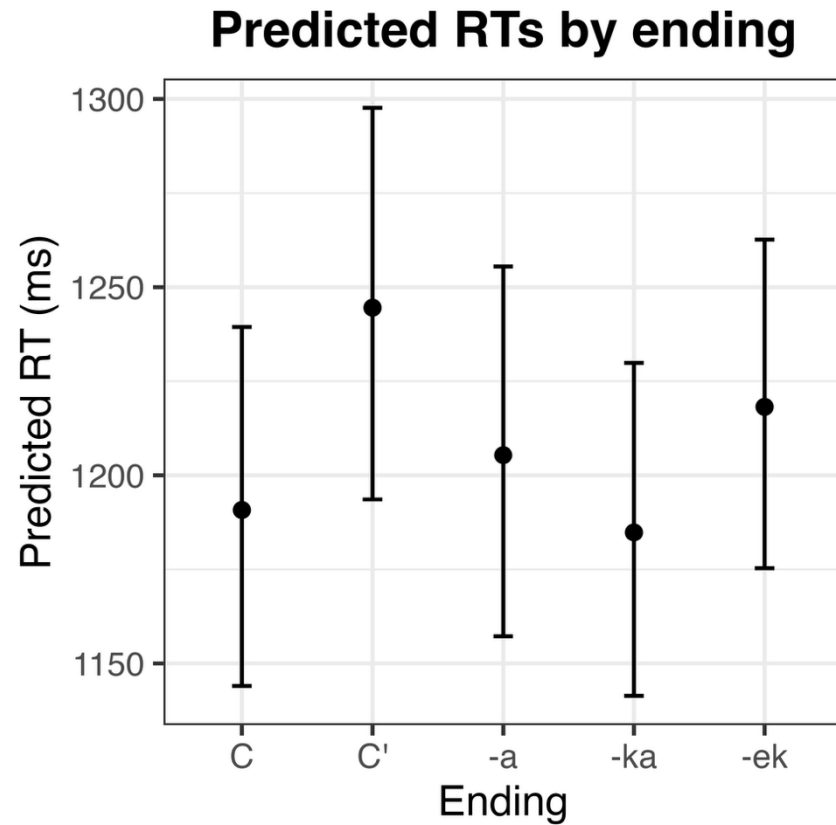
```
glmer(response_numeric ~ ending + (1 | part_id) + (1 + ending | base))
```

Exp 1: results



```
glmer(response_numeric ~ ending * consonant + (1 | part_id) + (1 + ending | base))
```

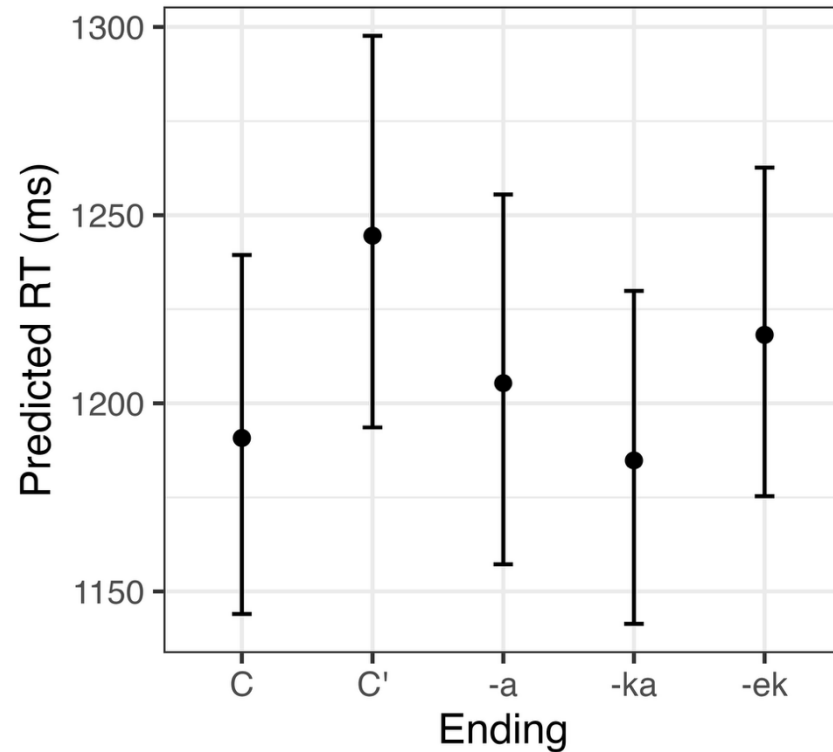
Exp 1: results



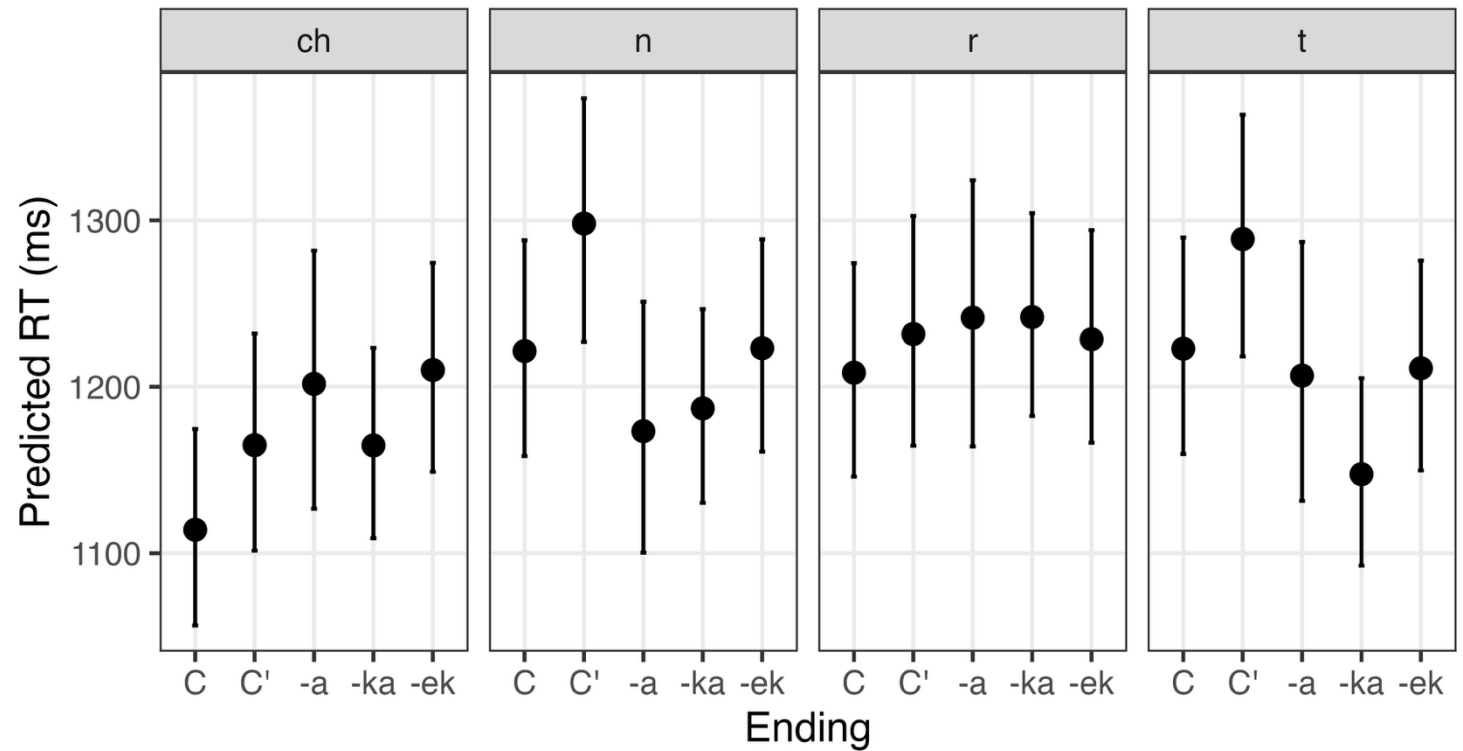
```
lmer(log(rt) ~ ending + (1 | part_id) + (1 + ending | base))
```

Exp 1: results

Predicted RTs by ending



Predicted RTs by consonant x ending



```
lmer(log(rt) ~ ending * consonant + (1 | part_id) + (1 + ending | base))
```

Exp 1: discussion

- Speakers relatively systematically assign nouns as masculine or feminine based on the final segments
- The level of systematicity seems reflect the systematicity in the existing phonological mapping
- Palatal consonants are a much less clear cue and are not a homogenous category

Exp 1: discussion

- Speakers relatively systematically assign nouns as masculine or feminine based on the final segments
- The level of systematicity reflects the systematicity in the existing phonological mapping
- Palatal consonants are a much less clear cue and are not a homogenous category

But what happens when we do not elicit gender assignment?

Experiment 2, 3 & 4

dyr
dyř
Zaznamenali jsme, že dyra je slepená a nejde rozlepit
dyrka slepený
dyrek

‘We noticed that [NONWORD] is glued together [f/m] and cannot be separated.’

INSTRUCTION: “made-up Czech-like words referring to inanimate entities”

Experiment 2, 3 & 4

Zaznamenal jsem, že

dyr
dyř
dyra *je* *slepená*
dyrka *slepený* *a nejde rozlepit*
dyrek

Zaznamenal(i) _____.

Experiment 2, 3 & 4

Zaznamenali jsme, že *dyr* *dyř* *dyra* *je* *slepená* *a nejde rozlepit*
dyrka
dyrek *slepený*

_____ jsme, _____.

Experiment 2, 3 & 4

Zaznamenali jsme, že *dyr* *dyř* *je* *slepená* *a nejde rozlepit*
dyra *slepený*
dyrka
dyrek

_____ že _____.

Experiment 2, 3 & 4

Zaznamenali jsme, že *dyr*
 dyř
 dyra *je* *slepená* *a nejde rozlepit*
 dyrka *slepený*
 dyrek

_____ *dyr* _____.

Experiment 2, 3 & 4

Zaznamenalí jsme, že *dyr* *dyř* *je* *slepená* *a nejde rozlepit*
dyra
dyrka
dyrek *slepený*

_____ je _____.

Experiment 2, 3 & 4

Zaznamenali jsme, že *dyr*
 dyř
 dyra *je* *slepená* *a nejde rozlepit*
 dyrka *slepený*
 dyrek

_____ slepená _____.

Experiment 2, 3 & 4

Zaznamenali jsme, že *dyr* *dyř* *dyra* *je* *slepená* *a nejde rozlepit*
dyrka
dyrek *slepený*

_____ a _____.

Experiment 2, 3 & 4

Zaznamenali jsme, že *dyr*
 dyř
 dyra *je* *slepená* *a nejde rozlepit*
 dyrka *slepený*
 dyrek

_____ nejde _____.

Experiment 2, 3 & 4

Zaznamenali jsme, že *dyr* *dyř* *dyra* *je* *slepená* *a nejde rozlepit*
dyrka
dyrek *slepený*

_____ rozlepit.

Exp 2, 3 & 4: structure & analysis

Experiment	Contrast	Example	N _{targets} : N _{fillers}	N _{participants}
2	C x -a	čut x čuta	20 : 80	305
3	-ka x -ek	čutka x čutek	20 : 80	286
4	C x C'	čut x čuť	20 : 80	318

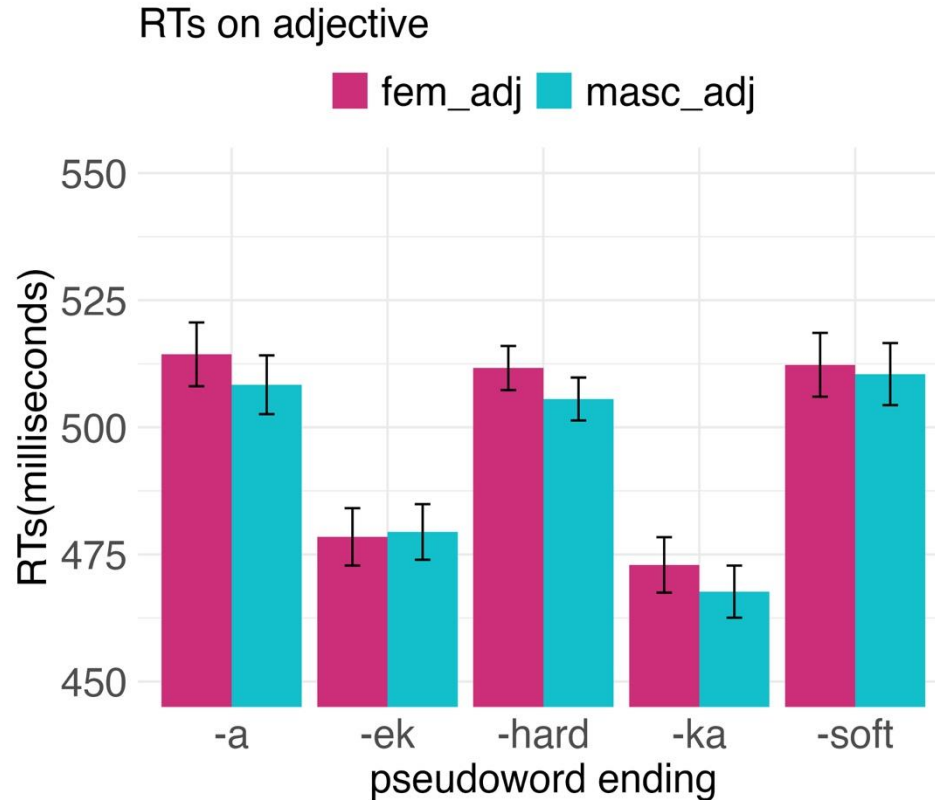
Exp 2, 3 & 4: structure & analysis

Experiment	Contrast	Example	N _{targets} : N _{fillers}	N _{participants}
2	C x -a	čut x čuta	20 : 80	305
3	-ka x -ek	čutka x čutek	20 : 80	286
4	C x C'	čut x čuť	20 : 80	318

- Analyzed all together
- Predicting log-transformed reading times on the adjective and the spillover region (two following words)

e.g., Zaznamenali jsme, že dyra je slepený a nejde rozlepit.

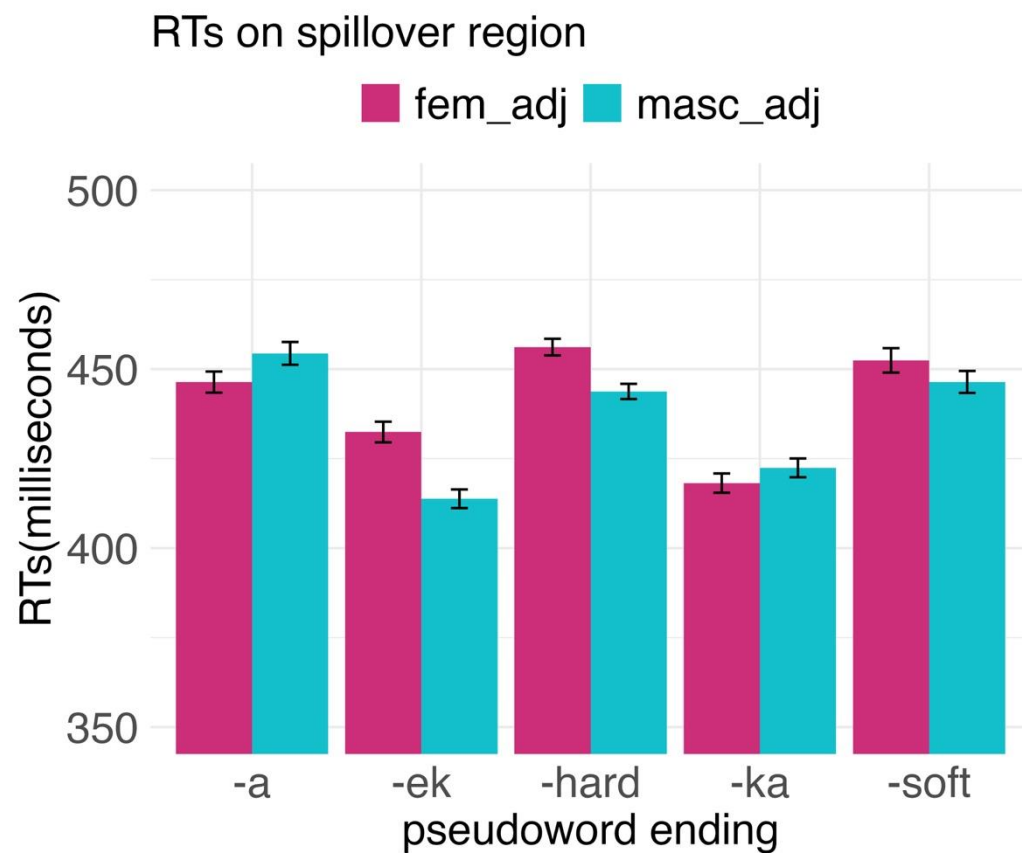
Exp 2, 3 & 4: results



<i>Predictors</i>	<i>log(rt)</i>		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	6.16	6.12 – 6.21	<0.001
final cons [-ek]	-0.07	-0.12 – -0.02	0.010
final cons [-hard]	-0.00	-0.03 – 0.03	0.944
final cons [-ka]	-0.07	-0.13 – -0.02	0.005
final cons [-soft]	0.01	-0.04 – 0.06	0.644
agr [m]	0.00	-0.02 – 0.02	0.857
frequency	-0.00	-0.00 – -0.00	0.012
final cons [-ek] × agr [m]	0.00	-0.03 – 0.03	0.817
final cons [-hard] × agr [m]	-0.01	-0.03 – 0.02	0.592
final cons [-ka] × agr [m]	-0.01	-0.04 – 0.02	0.485
final cons [-soft] × agr [m]	-0.00	-0.03 – 0.03	0.874

```
lmer(log(rt) ~ final_cons * agr + frequency + (1|base) + (1+final_cons|record))
```

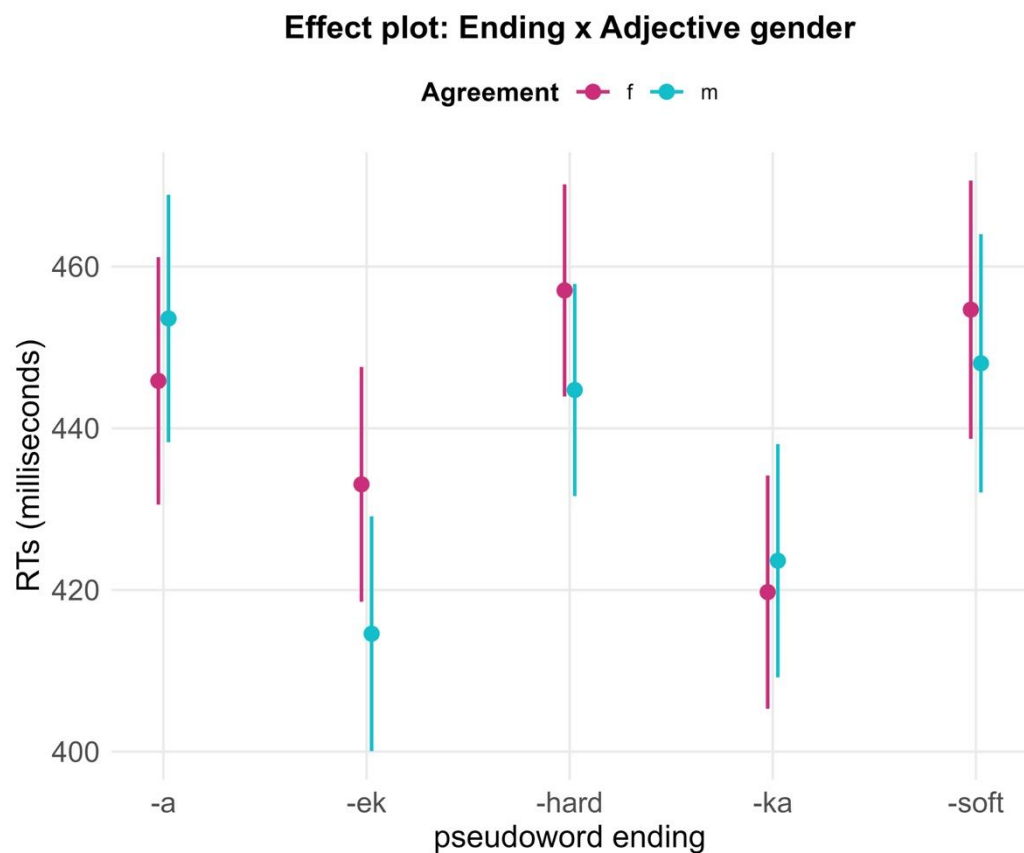
Exp 2, 3 & 4: results



<i>Predictors</i>	<i>log(rt)</i>		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	6.05	6.01 – 6.08	<0.001
final cons [-ek]	-0.03	-0.07 – 0.00	0.053
final cons [-hard]	0.02	-0.00 – 0.04	0.089
final cons [-ka]	-0.06	-0.10 – -0.03	<0.001
final cons [-soft]	0.01	-0.03 – 0.04	0.631
agr [m]	0.01	-0.00 – 0.03	0.060
final cons [-ek] × agr [m]	-0.05	-0.07 – -0.03	<0.001
final cons [-hard] × agr [m]	-0.03	-0.05 – -0.02	<0.001
final cons [-ka] × agr [m]	-0.00	-0.02 – 0.02	0.906
final cons [-soft] × agr [m]	-0.02	-0.04 – -0.00	0.039

```
lmer(log(rt) ~ final_cons * agr + frequency + (1|base) + (1+final_cons|record))
```

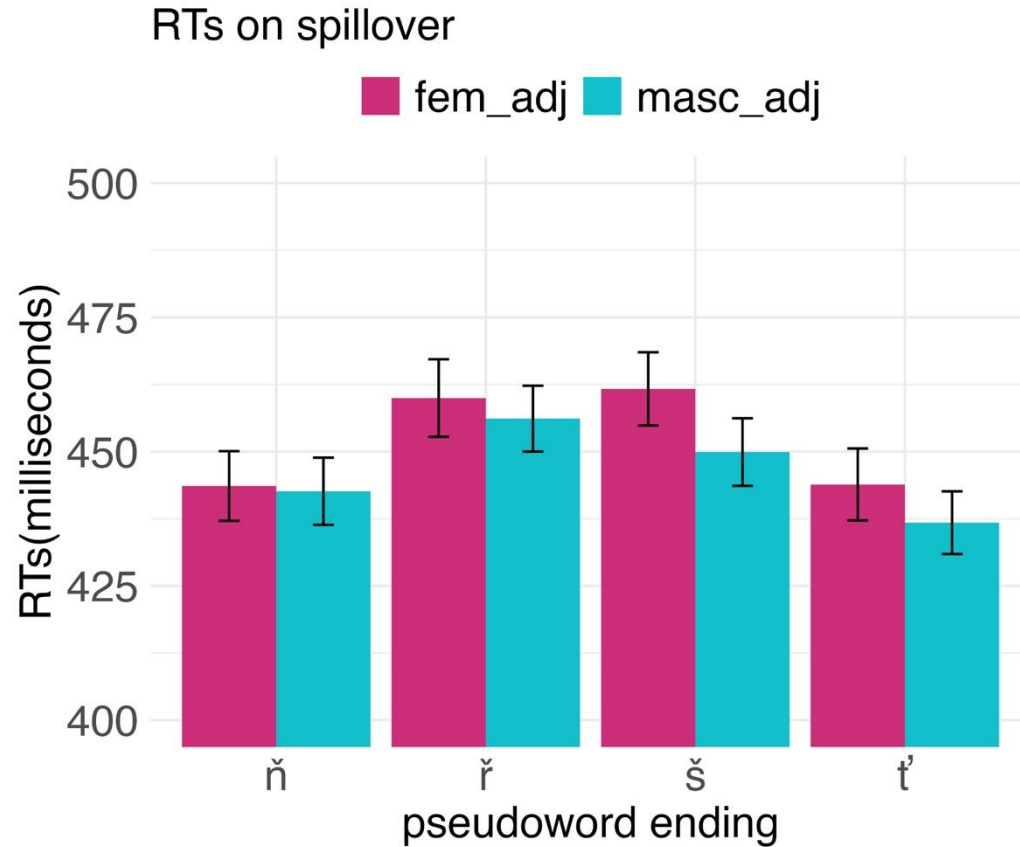
Exp 2, 3 & 4: results



Predictors	log(rt)		
	Estimates	CI	p
(Intercept)	6.05	6.01 – 6.08	<0.001
final cons [-ek]	-0.03	-0.07 – 0.00	0.053
final cons [-hard]	0.02	-0.00 – 0.04	0.089
final cons [-ka]	-0.06	-0.10 – -0.03	<0.001
final cons [-soft]	0.01	-0.03 – 0.04	0.631
agr [m]	0.01	-0.00 – 0.03	0.060
final cons [-ek] × agr [m]	-0.05	-0.07 – -0.03	<0.001
final cons [-hard] × agr [m]	-0.03	-0.05 – -0.02	<0.001
final cons [-ka] × agr [m]	-0.00	-0.02 – 0.02	0.906
final cons [-soft] × agr [m]	-0.02	-0.04 – -0.00	0.039

```
lmer(log(rt) ~ final_cons * agr + frequency + (1|base) + (1+final_cons|record))
```

Exp 4: results (palatals)



Predictors	log(rt)		
	Estimates	CI	p
(Intercept)	6.05	5.98 – 6.12	<0.001
final [ř]	0.02	-0.07 – 0.11	0.682
final [š]	0.03	-0.06 – 0.12	0.501
final [ť]	-0.00	-0.09 – 0.09	0.914
agr [m]	0.00	-0.07 – 0.08	0.946
frequency	-0.00	-0.00 – 0.00	0.212
final [ř] × agr [m]	-0.00	-0.13 – 0.13	0.983
final [š] × agr [m]	-0.02	-0.13 – 0.08	0.684
final [ť] × agr [m]	-0.01	-0.14 – 0.12	0.858

```
lmer(log(rt) ~ final_cons * agr + frequency + (1|base) + (1|record))
```

Experiment 2, 3 & 4: discussion

When encountering novel words in a sentential context, speakers...

... form expectations about their grammatical gender based on their final segment

... but only in the prototypical cases (–ek, C)

... and not when the final segment is more ambiguous (–a, –ka, C')

... animacy does not seem to play a role (in the early stage of processing)

General discussion

- Czech speakers relatively systematically assign novel nouns into grammatical gender categories based on their final segments
- Their level of systematicity in the assignment seems to mimic the level of systematicity in the grammatical gender system
- What other formal properties play a role?
 - N-grams, co-occurrence statistics, vector space models (Bojanowski et al, 2017)
- What about non-nominative forms?

Thank you!

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5, 135-146.

Corbett, G. G. (1991). *Gender*. Cambridge University Press.

Bull, W. (1965): Spanish for teachers: Applied Linguistics. New York: The Ronald Press Co

Szagun, G., Stumper, B., Sondag, N., & Franik, M. (2007). The acquisition of gender marking by young German-speaking children: Evidence for learning guided by phonological regularities. *Journal of Child Language*, 34(3), 445-471.



Co-funded by
the European Union



Ministry of Education,
Youth and Sports
of the Czech Republic



The work on this presentation was supported by
„Mediated Society (MEDIS:ON)
CZ.02.01.01/00/23_025/0008713“ which is co-
financed by the European Union.

<https://earlygrammarlab.github.io/people.html>