

Sublexical and distributional cues for grammatical gender agreement: evidence from self-paced reading times on Czech pseudowords.

Martina Abbondanza (abbondanza@praha.psu.cas.cz)¹, Mikuláš Preininger¹, Lukáš Jiříčka¹,
Filip Smolík¹

¹Institute of Psychology, Czech Academy of Sciences, Prague

Background. Czech exhibits a richly inflected grammatical gender (GG) system with four categories: masculine animate/inanimate, feminine, and neuter. Phonological and morphological cues to GG are often inconsistent with frequent cross-class syncretism [1]. Nevertheless, general regularities linking words to GG exist, and children exploit gender features for predictive processing [2]. Crucially, gender categorization may also be shaped by distributional information reflecting learned regularities. **This study** investigates the formal cues Czech speakers use for grammatical gender categorization testing pseudowords in self-paced reading tasks, with a particular focus on sub-lexical markers and distributional semantic information as predictors of reading performance. **Materials.** Pseudowords were constructed from 20 Czech-like pseudo-stems combined with five endings: non-palatalized (“hard”) consonants (e.g., *dyr*), palatalized (“soft”) variants (e.g., *dyř*), *–a* (e.g., *dyra*), *–ka* (e.g., *dyrka*), and *–ek* (e.g., *dyrek*). These pseudowords were embedded in subject-position sentences (Table 1) and followed by gender-marked adjectives (masculine or feminine). **Methods.** Three self-paced reading experiments (Exp. 1, N = 305; Exp. 2, N = 286; Exp. 3, N = 318) used a moving-window paradigm, each targeting a different contrast: *–a* vs. hard, *–ek* vs. *–ka*, and soft vs. hard. Lists were counterbalanced, so each participant saw each pseudoword in only one gender condition. A continuous gender index was derived from Czech FastText embeddings [3]. For each pseudoword, cosine distances were computed with respect to a set of masculine and feminine adjective anchors. Distances to each set were averaged, and the gender index was calculated as the difference between the mean distance to masculine anchors and the mean distance to feminine anchors ($\text{mean}_{\text{masc}} - \text{mean}_{\text{fem}}$). Lower values of the index indicate that a pseudoword is positioned closer to feminine anchors in the vector space, whereas higher values indicate greater proximity to masculine anchors. The gender index was included as a predictor in RT models. **Results.** Data were combined across experiments. A LMM with final segment, agreement, and their interaction (plus random slopes for participants and intercepts for participants and items) showed a significant interaction in the spillover region (Fig. 1), indicating reliance on morpho-phonological cues for *–ek*, and hard endings. A second model showed that the gender index significantly modulated RTs in the spillover region ($t = -2.97$, $p = .003$; Fig. 2). **Discussion.** Results show that Czech speakers rely on sub-lexical morpho-phonological cues to guide grammatical gender agreement, but

that these cues vary significantly in their predictive strength. Clear effects for –ek and hard endings suggest that speakers draw on them robust form–gender regularities, whereas the lack of effect for –a, –ka and soft endings reflect lower reliability of these cues in the input. Crucially, the results are confirmed by the distributional semantic analysis as the interaction between the gender index extracted from FastText and the gender of the adjective was particularly evident as the index was higher (i.e. more masculine). This pattern is consistent with a usage-based and probabilistic account of grammatical gender, in which local morpho-phonological cues are integrated with distributional experience accumulated from the lexicon. [4, 5].

References

- [1] Naughton, J., & Von Kunes, K. (2020). *Czech: An essential grammar*. Routledge.
- [2] Smolík, F., & Bláhová, V. (2019). Czech 23-month-olds use gender agreement to anticipate upcoming nouns. *Journal of Experimental Child Psychology*, 178, 251-265.
- [3] 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.
- [4] Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in cognitive sciences*, 10(7), 335-344.
- [5] Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., & Tenenbaum, J. B. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*.

Table 1. Example of the experimental sentences used in the self-paced reading tasks.

Region	Text	Gloss
intro	Zaznamenali jsme, že	<i>we noticed that</i>
pseudoword	dyr [hard] / dyř [soft] / dyra [-a] / dyrek [-ek] / dyrka [-ka]	<i>pseudoword</i>
copula	je	<i>is</i>
adjective	slepený [M] / slepená [F]	<i>stuck together</i>
spillover	a nejde	<i>and cannot</i>
final	rozlepit.	<i>be separated</i>

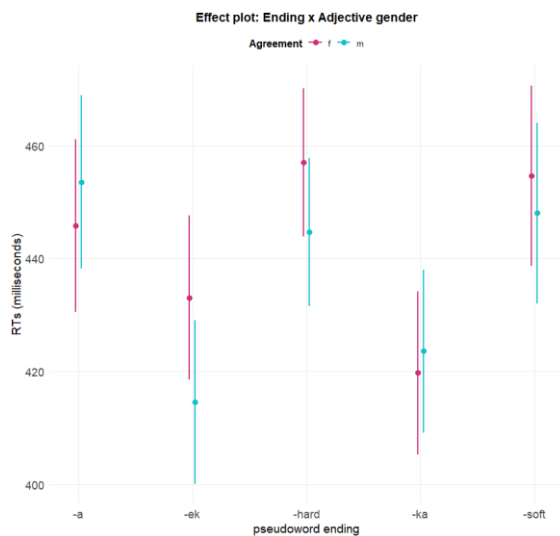


Figure 1. Effect plot of the linear-mixed effects model with Final segment (on the x-axis) X Agreement (i.e. adjective gender: red= feminine, blue= masculine).

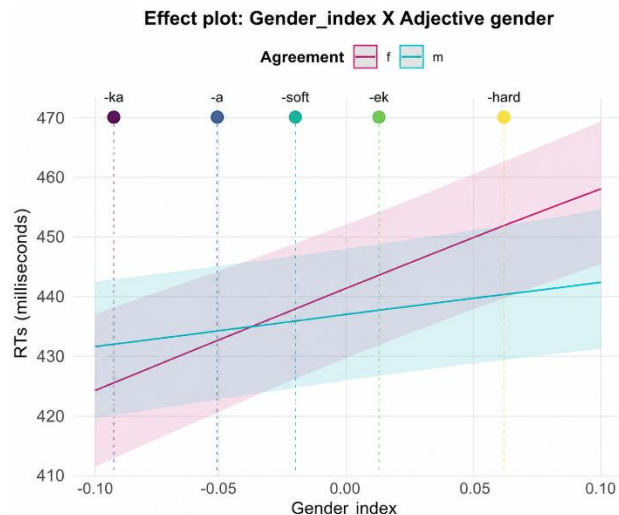


Figure 2. Effect plot of the linear-mixed effects model with Gender index (on the x-axis) X Agreement (i.e. adjective gender: red= feminine, blue= masculine). Mean indexes for the endings are shown on the top of the x-axis.

Czech nominal gender system

The Czech grammatical gender (GG) system comprises four categories—masculine animate, masculine inanimate, feminine, and neuter—which primarily dictate adjective agreement and inflectional paradigms. The following table summarizes the general morpho-phonological regularities in the Czech gender system.

Grammatical gender	Typical final segments	Examples
Masculine in-animate	consonants	<i>hrad</i> (castle), <i>stroj</i> (machine)
Masculine animate	consonants, -a, -e	<i>pán</i> (mister), <i>muž</i> (man), <i>táta</i> (dad), <i>soudce</i> (judge)
Feminine	-a, -e, soft consonants, -st	<i>voda</i> (water), <i>růže</i> (rose), <i>píseň</i> (song), <i>radost</i> (joy)
Neuter	-o, -e, -í	<i>město</i> (town), <i>moře</i> (sea), <i>stavení</i> (building)