How Aunt-Like Are You? Exploring Gender Bias in the Genderless Estonian Language: A Case Study

Elisabeth Kaukonen[1] **Ahmed Sabir**[2] **Rajesh Sharma**[2]
[1]University of Tartu, Institute of Estonian and General Linguistics, Estonia
[2]University of Tartu, Institute of Computer Science, Estonia

Abstract

This paper examines gender bias in Estonian, a grammatically genderless Finno-Ugric language, which doesn't have gendered noun system nor any gendered pronouns, but expresses gender through vocabulary. In this work, we focus on the male-female compound words ending with *-tädi* 'aunt' and *-onu* 'uncle', aiming to pinpoint the occupations these words signify for women and men, and to examine whether they reveal occupational differentiation and gender stereotypes. The findings indicate that these compounds go beyond occupational titles and highlight prevalent gender bias.

1 Introduction

Languages are divided into three groups based on gender expression: firstly, there are grammatical gender languages (such as Russian, French, German, *etc*.), which use a gendered noun class system. Secondly, there are natural gender languages (*e.g*. English, Swedish, *etc*.), which incorporate gender-specific pronouns. Lastly, there are genderless languages (*e.g*. Hungarian, Finnish, Turkish, *etc*.), which lack gendered nouns as well as pronouns (Stahlberg et al., 2007). Estonian, representing a Balto-Finnic language, is grammatically genderless and thus incorporates only lexical resources, *i.e*. gender-specific vocabulary for gender expression.

While a grammatical gender does not correlate with gender equality or neutrality in a certain society (Aikhenvald, 2016), gender bias and stereotyping can still be prevalent not only in societies where a genderless language is spoken, but also within those languages themselves. This work illustrates how gender stereotypes are manifested in Estonian gendered vocabulary, specifically compound words ending with lemmas *tädi* and *onu* that refer to occupations, shedding light on which professions are more commonly associated with women or men, and thus, how a genderless language exhibits bias and stereotypes. The gender stereotypes referred to here are mainly beliefs about occupational and social roles that are assumed to be held by men or by women more dominantly (Gygax et al., 2016; Vaidya, 2021).

In Estonian, the terms *tädi* and *onu* are primarily used to denote kinship, however, they also serve other purposes. For instance, they are commonly used in children's language, when referring to unfamiliar individuals or family friends when talking to children. Additionally, *tädi* and *onu* can be used humorously and they frequently appear in compound words denoting occupations (Puna, 2006). Such words were chosen for this paper, since they represent more informal and non-standardized language use. Furthermore, as these words represent informal language, they might reflect stereotypes more directly and with less linguistic filtering, as opposed to potentially more moderated words used in formal contexts. Examining gender bias in genderless languages, such as Estonian, is crucial because this topic has received little attention in the context of low-resource languages. Such languages still provide valuable insights into gender dynamics and social beliefs, which help to identify harmful and discriminatory gender stereotypes as well as raise awareness of gender inequality and occupational segregation. The research questions this study aims to address are as follows: (1) What kinds of occupations do the compound words ending with *tädi* and *onu* express? (2) How do occupational titles ending with *tädi* (aunt) and *onu* (uncle) propagate gender bias in Large Language Models (LLMs)?

2 Gender Expression in Estonian

Gender in Estonian is only expressed through vocabulary. This can be done, for instance, by using

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 296–301
March 3-4, 2025 ©2025 University of Tartu Library

Male-dominated		Female-dominated	
occupation	%	occupation	%
Doctor	84	Cashier, shopkeeper	80
Construction worker	1	Cook	72
Security worker	22	Librarian	98
Bus or tram driver	10	Kindergarten teacher	99
Electrician	1	High school teacher	86
EU politician	27	Receptionist	74
IT support specialist	28	Ticket seller	91
Waste collector	0	Social worker	92
Warehouse worker	8	Cleaner	88
Mailman	40	Hairdresser	94

Table 1: The percentage of females in male- and female-dominated occupations (%) in the Estonian labor force statistics, 2021.

separate words (*e.g.* mees 'man', naine 'woman', tüdruk 'girl', poiss 'boy', ema 'mother', isa 'father'). In addition, another option to express gender is through compounding. This means adding two single words together, one of which carries a gendered meaning. There are two ways to indicate gender with a compound word in Estonian. Firstly, gendered prefixes *nais-* 'female' and *mees-* 'male' that function as adjectives can be added to a role noun (*e.g.* naisarst 'female doctor', naisujuja 'female swimmer', meesmodell). Secondly and similarly, gender-specific base forms (*i.e.* suffixes) can be used (*e.g.* esimees 'chairman', ärinaine 'businesswoman', spordinaine 'sportswoman'). In the second option, the noun indicating gender conveys the main meaning of the word.

There is also a third option - derivation, which specifically denotes female agents, including female representatives of different ethnicities, for example, lauljanna 'female singer', venelanna 'female Russian', poetess 'female poet', sõbratar 'female friend', 'girlfriend' *etc*. (Kasik, 2015; Hasselblatt, 2015). However, derivation, compared to single words and compound words, is perhaps not that widely used in everyday language today. Derivation is the only instance where a gendered morpheme is used in Estonian vocabulary.

As for compound words with a gendered base form, generally the most common nouns used in such compounds are *-mees* 'man' and *-naine* 'woman' (especially *-mees*, since *mees*-ending compounds are used generically, like esimees 'chairman'). However, other nouns, such as *-tädi* 'aunt' and *-onu* 'uncle' can also be included. (Gu, 1990; Clyne et al., 2009; Kiss, 2022).

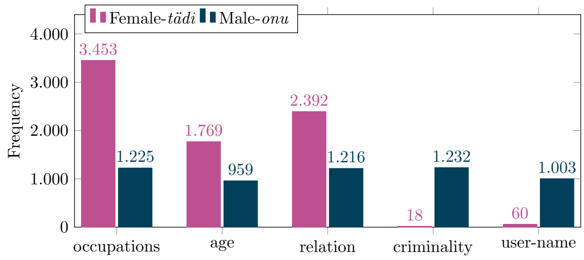

Figure 1: Frequencies of the more dominant semantic categories of gendered *tädi* and *onu* compound words that occurred in the Estonian web corpus.

3 Dataset

The dataset used in this study comes from the web subcorpus of the Estonian National Corpus of 2021, contains 724 million words (882 million tokens), with a variety of genres (*e.g.* online forums, e-commerce, online periodicals, property portals, recipe collections, *etc*.). To navigate the corpus, the SketchEngine tools (Kilgarriff et al., 2004) are used to extract compound words ending with lemmas *-tädi* and *-onu*. The extract token frequencies of compound words are 6500 (830 types) for the male compounds *-onu* and 6100 (700 types) for the female compounds *-tädi*. Compound words that occurred in the data were classified into semantic categories (see Figure 1), based on the meaning of the first part (or the prefix) of the compound. From these categories, words referring to occupations were specifically selected and chosen for analysis. The total number of occupational titles after preprocessing is 206 titles. We use the Estonian labor force statistics[1] to illustrate descriptive gender bias as shown in Table 1, the percentage distribution of females in male-dominated and female-dominated occupations.

4 Data Analysis and Result

4.1 Statistical Evaluation

As for words denoting occupations and activities, primary focus was on identifying the occupations associated with *tädi* and *onu* and whether titles denoting women and men correspond to different occupations, thereby revealing gender-based occupational stereotypes. To categorize occupational titles, words that appeared at least three times were considered. Table 2 shows an overview of the different types of occupations that emerged with *tädi*- and *onu*-compounds. The percentages show the

[1] https://palgad.stat.ee/

Occupation	Female-*tädi*-compounds			Male-*onu*-compounds			Examples
	occ	%	type	occ	%	type	
Customer service	1502	44	57	87	7	12	*raamatukogutädi* (library aunt), *garderoobitädi* (wardrobe aunt)
Healthcare	464	13	10	150	12	1	*arstionu* (doctor uncle), *haigladtädi* (hospital aunt)
social work	378	11	15	–	–	–	*koolitädi* (school aunt), *kasvatajatädi* (kindergarten teacher aunt)
Construction	–	–	–	66	5	8	*remondionu* (repair uncle), *toruonu* (pipe uncle)
Entertainment	42	1	8	58	5	7	*kunstitädi* (art aunt), *kaameraonu* (camera uncle)
Law-enforc	134	4	7	247	20	10	*turvatädi* (security aunt), *valvurionu* (guard uncle)
Journalism	29	1	4	35	3	4	*raadiotädi* (radio aunt), *leheonu* (newspaper uncle)
Business	–	–	–	14	1	4	*naftaonu* (petroleum uncle), *korp-onu* (corporate uncle)
Science	–	–	–	17	1	5	*teadlaseonu* (scientist uncle), *tehnikaonu* (technology uncle)
Politics	43	1	6	66	5	8	*riigonu* (government uncle), *eurotädi* (European parliament aunt)
Cleaning	130	4	3	7	1	1	*koristajatädi* (cleaning aunt), *prügionu* (garbage uncle)
Animal	111	3	8	4	0.3	1	*koeratädi* (dog aunt), *farmitädi* (farm aunt)

Table 2: Groups of occupations emerged occupational titles ending with female compounds *tädi* (aunt) and male compounds *onu* (uncle) expressed, including occurrences, percentages from the whole group, type frequencies and example words. Type frequency denotes the number of different compounds in the corpus (*i.e.* how many different *tädi*-compounds emerged).

proportion of certain types of occupations among all occupational title ending with either *tädi* or *onu*.

Tädi in occupational titles primarily marked professions related to customer service (44% from all occupational titles), healthcare (13%), and social work (11%), while *onu* in occupational titles predominantly represented law enforcement (20%), followed by healthcare (12%) and customer service (7%). Thus, women are more often associated with occupations related to children, teaching, and (elder) care, while men are often found in the role of guards and police officers. As for *tädi*-compounds, there were no instances of words expressing occupations related to repairing and construction, business and entrepreneurship, and science and technology. Therefore, occupational titles ending with *tädi* and *onu* reflect the traditional gender associations regarding occupations, highlighting those typically attributed to women and men (Kaukonen, 2023).

4.2 LLMs Evaluation

In this section, we examine the propagation of occupational title biases in compound words ending with *tädi* (female 'aunt') and *onu* (male 'uncle') in LLMs. For this, inspired by the human-written CrowS-Pairs dataset (Nangia et al., 2020), which uses sentence pairs to highlight stereotypes across social categories, we manually created sentence pairs using the same Estonian National Corpus (see Section 3). These pairs are based on 87 occupations, with one occupation per pair of sentences (in total 174 sentences) where the occupational bias can be used with either gendered compound word (see Table 3) *e.g.* "The [cleaning aunt/uncle] carefully dusted the drawers.". The Estonian labor force statistics database (2021) is also used as a reference to identify descriptive gender bias, reflecting gender-stereotyped professions.

We employ the most recent state-of-the-art LLMs models, ChatGPT (OpenAI, 2022), GPT-4 (Achiam, 2023), GPT-4-Turbo, GPT-4o (OpenAI, 2024a), GPT-o1 (OpenAI, 2024b), LLAMA-3 (Touvron et al., 2023) (8B and 70B models), and LLAMA-2-7B fine-tuned Estonian models LLAMMAS (Kuulmets et al., 2024): (1) LLAMMAS-base that is fine-tuned on 5B tokens (Both are fine-tuned on 75% Estonain, 25% English dataset), (2) LLAMAAS is an improved version of LLAMAAS-base that is additionally instruction-tuned, and (3) LLAMAAS-MT is additionally finetuned on translation instructions on English-Estonian dataset.

For the prompt-based model, we set the temperature parameter to zero through all experiments to ensure consistent output and run the experiments three times. Majority voting is used to finalize the model's decision, except for the reasoning-based GPT-o1 model. For the LLAMA models (LLAMA-3 and LLAMMAS), we extract the mean probability of all tokens in the sentences containing compound words as the occupational title gender bias score, as shown in Table 3 with LLAMA models.

Gender Bias Amplification Score. For evaluation, we measure the bias amplification as correlation measure (Zhao et al., 2017) towards the protected attribute $g \in$ *uncle* male or *aunt* female compounds words and the occupational title:

$$b(\text{occ}, g) = \frac{c(\text{occ}, g)}{\sum_{g' \in \{\text{m,f}\}} c(\text{occ}, g')}$$

where $c(\text{occ}, g)$ is the occurrences of the occupa-

Model	Sentence with Compound Words English Translation with Original Estonian Sentence	Bias Ratio to-Uncle	Bias Ratio to-Aunt
Eng	Baker aunt/Baker uncle made delicious pretzels		
LLAMMAS	pagari*tädi*d/pagari*onu*d valmistasid maitsvaid kringleid	0.48	0.52
LLAMA-3-70B	pagari*tädi*d / pagari*onu*d valmistasid maitsvaid kringleid	0.46	0.54
Eng	The cleaning aunt/cleaning uncle carefully dusted the drawers		
LLAMMAS	Koristaja*onu* /Koristaja*tädi* pühkis hoolega tolmu kummutilrecip	0.48	0.51
LLAMA-3-70B	Koristaja*onu* /Koristaja*tädi* pühkis hoolega tolmu kummutilrecip	0.50	0.49

Table 3: Examples of occupational title bias using the fine-tuned Estonian LLAMA (LLAMMAS) and the off-the-shelf LLAMA-3-70B models. (Top) The example demonstrates how the models measure gender bias, associating bakery tasks with women. (Bottom) In the example with the cleaning [aunt/uncle] occupational title, the standard LLAMA-70B incorrectly reflects the female-biased occupation.

Model	Occupational Title Ratio M *onu*	Occupational Title Ratio F *tädi*	Occupational Title Ratio %
Labor Force Data	0.37	0.63	
ChatGPT (OpenAI, 2022)	0.43	0.57	0.64
GPT-4 (Achiam, 2023)	0.68	0.32	0.66
GPT-4-Turbo	0.63	0.37	0.71
GPT-4o (OpenAI, 2024a)	0.34	0.66	0.83
GPT-o1 (OpenAI, 2024b)	**0.36**	**0.64**	**0.85**
LLAMA-3-8B (Touvron et al., 2023)	0.46	0.54	0.52
LLAMA-3-70B	0.38	0.62	0.60
LLAMMAS (Kuulmets et al., 2024)	0.47	0.53	0.64
LLAMMAS-Base	0.48	0.52	0.63
LLAMMAS-MT	0.55	0.45	0.49

Table 4: Comparison result between different LLMs on occupational title using *tädi* and *onu* compound word. For the LLAMA-3 and Estonian LLAMMAS-7B, we rely on the mean probability, of the sentence with the bias occupations, for measuring the bias. The results indicate that the GPT-o1 model aligns closely with Estonian labor force statistics.

tions and the male-female compound words ending with *tädi* and *onu*. Table 4 shows a comparison results between different state-of-the-art LLMs. The best model aligned with labor force statistics is GPT-o1, especially concerning less common biased occupational titles (*e.g.* piima*tädi*, which refers to milk lady). The GPT-4o model achieved a comparable alignment level of 83%. The Estonian fine-tuned model LLAMA-2-7B (LLAMMAS) reflects the biases more accurately than the standard LLAMA-3 models with a 4-point difference in descriptive bias alignment compared to the 70B model.

Table 3 shows examples of the open-source model bias scores for the fine-tuned model LLAMMAS and the standard LLAMA-3-70B. The bottom example shows that the off-the-shelf model incorrectly reflects a female-biased occupational title from the labor data, *cleaning aunt/uncle*.

5 Discussion

The analysis of compound words suggests women typically assume caregiving roles and are often associated with children, while men occupy professions like law enforcement. Additionally, men are more common in fields such as construction, business, entrepreneurship, and science. Conversely, the data indicates that men are rarely seen working in the educational sector. While this could indicate coincidental occupational gender differences, the results appear to reflect sectorial segregation, with women overrepresented in low-paid sectors like care, education, and customer service. Evidence from the 2021 Estonian Census supports this, showing 86% of healthcare and social welfare, 83% of education, and 82.3% of the service sector workers are women.

The analysis of LLMs revealed that these models propagate occupational biases related to compound words. Specifically, the fine-tuned Estonian LLAMMAS model reflects biases from Estonian labor force statistics more accurately than the similar-sized LLAMA-3-8B and larger LLAMA-3-70B models. This indicates that the process of fine-tuning has amplified the inherent biases within the model. For instance, the secretary as a female-biased occupation aligned correctly with all models (except for LLAMA family). However, in the fine-tuned model that incorporates additional parallel data (English-Estonian sentence pairs), the labor force data alignment bias ratio is lower compared to all other models, particularly for highly female-biased occupations (*e.g.* nanny, hairdresser, *etc.*).

6 Conclusion

This paper examined Estonian compound words ending with *tädi* (‘aunt’) and *onu* (‘uncle’) in the Estonian web corpus 2021. The findings indicate these terms reflect traditional gender roles and stereotypes in occupational contexts, which are also mirrored by LLMs, reinforcing gender biases.

Limitation

The limitations of the present study include the analysis of only informal gendered language units such as compounds ending with *tädi* and *onu*. If, for example, *-mees* ’man’ and *-naine* ’woman’ compounds, some of which constitute official occupational titles, were examined, then a more broad view of entrenched stereotypes could be achieved. Furthermore, several of the examined occupational titles were low in frequency as well as expressing quite novel or uncommon professions. As for the analysis of usage, the study included only specific uses of *tädi* and *onu*, and such an analysis may not translate to all other cases.

Ethics Statement

In this work, we measure gender bias patterns using descriptive modeling, which reflects observed real-world statistics. However, we also recognize the importance of normative analysis, which provides critical insights into promoting fairness and achieving equitable and unbiased outcomes. Balancing these approaches contributes to building a more just and inclusive society.

The corpus used in this study have been obtained from publicly available sources and have been anonymized. Any conflicts of interest or biases that may influence the interpretation of results are acknowledged. The authors acknowledge that this approach to gender does not encompass the entirety of gender identities, many of which are not represented by this vocabulary. Furthermore, only one bias considering gender is addressed in this paper, while the dataset may contain other demographic biases, such as race, religion and nationality. Also, this study focuses on occupational titles ending with *tädi* (aunt) and *onu* (uncle), which may propagate specific gender biases tied to cultural stereotypes regarding roles traditionally associated with women or men.

Acknowledgment

This work has received funding from the EU H2020 program under the SoBigData++ project (grant agreement No. 871042), by the CHIST-ERA grant No. CHIST-ERA-19-XAI-010, (ETAg grant No. SLTAT21096), and partially funded by HAMISON project.

References

Josh et al. Achiam. 2023. Gpt-4 technical report. *arXiv:2303.08774*.

Alexandra Y Aikhenvald. 2016. *How gender shapes the world*. Oxford University Press.

Michael G Clyne, Catrin Norrby, and Jane Warren. 2009. *Language and human relations: Styles of address in contemporary language*. Cambridge University Press.

Yueguo Gu. 1990. Politeness phenomena in modern chinese. *Journal of pragmatics*.

Pascal M. Gygax, Alan Garnham, and Sam Doehren. 2016. What do true gender ratios and stereotype norms really tell us? *Frontiers in Psychology*, 7.

Cornelius Hasselblatt. 2015. The representation of gender in estonian. *Gender Across Languages. The linguistic representation of women and men*, 4:125–151.

Reet Kasik. 2015. *Sõnamoodustis*. Tartu Ülikooli Kirjastus.

Elisabeth Kaukonen. 2023. Cleaning aunts and police uncles in action. unveiling gender dynamics in estonian compound words. *Journal of Estonian and Finno-Ugric Linguistics*, 14(3):137–171.

Adam Kilgarriff, P Rychly, P Smrz, D Tugwell, G Williams, and S Vessier. 2004. The sketch engine. pages 105–115.

Orsolya Kiss. 2022. Forms of address in the tatar language spoken in finland and estonia. Master’s thesis, University of Tartu, Tartu.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. 2024. Teaching llama a new language through cross-lingual knowledge transfer.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *EMNLP*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2024a. Gpt-4o system card. `https://openai.com/index/gpt-4o-system-card/`. Accessed: 2024-08-08.

OpenAI. 2024b. Openai o1 system card.

Kerli Puna. 2006. *Soospetsiifilised isikunimetused sõnaraamatutes ja tekstides*. Ph.D. thesis, Tartu Ülikool, Tartu.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Shubhangi Vaidya. 2021. *Gender Stereotypes*, page 654–663.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.