

Occupational gender bias in ungendered languages: Comparing experimental data from Hungarian and Chinese

Anonymous ACL submission

Abstract

This paper is about occupational gender bias and stereotypes, presented in a cross-cultural setting. In the study, we analyze experimental data collected from Hungarian and Chinese speakers on their ratings of occupational titles, answering a question on how typically a job is done by either men or women. Results show that in both of these languages the words carry societal biases, despite that the job titles themselves have no gender markings. We compare the ratings across linguistic and participant gender lines, highlight the differences, and discuss the results with insights ranging from peculiarities in word formation to more generic societal differences. We also compared the human raters' responses with that of a few popular generative AI engines, which will show that the biases we humans carry are even stronger in the Large Language Models (LLMs) underlying these chatbots.

1 Introduction

2 Experiment setup

For both languages we devised a simple experiment in survey form, where we asked participants to rate job titles on a 7-point Likert scale. In both cases they were instructed to make decisions on how likely that occupation is to be done by men or by women, according to their own perception. The exact wording in translation were: HU: "Is the occupation typically a man's occupation or a woman's occupation?"; ZH: "What do you think is the ratio of men to women in `occupation_name`?". The scale presented in both questionnaires followed the same logic, from -3 to +3, moving from men to women, with 0 in the middle, hence the choices were completely male; mostly male; somewhat male; neutral/equal; somewhat female; mostly

female; completely female. First, we will introduce the Hungarian survey, then the Chinese one, and finally we will compare the results of the two.

2.1 Hungarian survey

2.1.1 Materials

The Hungarian survey contained 50 items, each a commonly occurring job title in Hungary, such as: *modell* 'model' or *katona* 'soldier', in no particular order. Six of the words were attention-check items, which were removed from the final analysis. The attention checks were *pincérnő* 'waitress', *titkárnő* 'secretary (female)', *tanárnő* 'teacher (female)', *takarítónő* 'cleaning lady', *ápolónő* 'nurse (female)', and *háztvezetőnő* 'housekeeper (female)'. These words explicitly determine the gender of the worker by appending *-nő* 'woman' to the base word. If participants paid attention, all these items had to be rated according to 'completely female'. Participants who failed even one, were rejected.

These words above have also have their counterparts without the suffix, *pincér* 'waiter', *titkár* 'secretary', *tanár* 'teacher', etc., these are unmarked for gender. Common pairs include *énekes* 'singer' – *énekesnő* 'female singer', *színész* 'actor' – *színésznő* 'actress', and in such cases where both are well established, the unmarked word seems to have some male bias, but it does not explicitly refer to a man. More interestingly, there are occupations where the unmarked form is the only one generally used for both genders, and appending *-nő* 'woman' to it – although possible – would render it a bit awkward, such as in *alkalmazottnő* 'female employee' or *programozónő* 'female programmer', but not impossible such as English *singress* would be. Furthermore, there are a

few cases, where the female-marked version is so widespread, that it is the unmarked version that will sound odd, such as *házvezető* ‘housekeeper’, or to some extent *takarító* ‘cleaner’.

In short, we are interested in these unmarked words, as they do not inherently carry a male bias, but according to our expectations will nonetheless be rated according to the prevailing societal stereotypes.

We also included *diák* ‘student’ out of curiosity. Although being a student is not a job, but it is beyond doubt the only truly gender-neutral “occupation” there is, since it is mandatory for every child to go to school both in Hungary and in China. We wanted to see if there would be any bias regarding this word, especially that Hungarian has a female form for it, *diáklány* ‘girl student’.

2.1.2 Procedure

The questionnaire was distributed online, and after a brief welcome message and the instructions, the words were presented in a simple list format, each word with a corresponding rating scale next to it, with no context. Time limit was not set, but the survey was designed to take around 5 minutes; participants took 4 minutes 25 seconds on average to finish.

2.1.3 Participants

A total of 21 native Hungarian speakers filled the questionnaire, and after validating the responses (reviewing attention checks and manual checking for anomalies) 2 were rejected. The participants were mostly recruited through Prolific, an online platform, with screeners for current location (Hungary) and first language (Hungarian). Participants were compensated for their time with a small reward. In the end, the Hungarian rating dataset had 20 participants (11 female, 9 male), with ages ranges of 25-35 (n=11), 35-45 (n=4), and 45-55 (n=5). See Figure 1 for the distribution.

3 Mandarin Chinese survey

4 Results

...overall

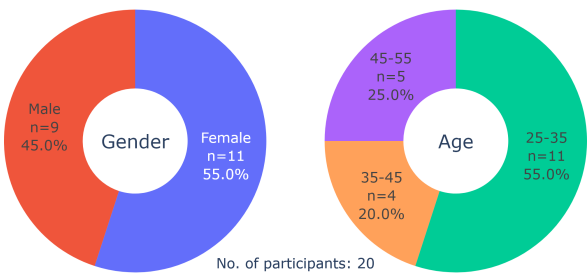


Figure 1: Demographics of the Hungarian participants.

4.1 Hungarian	124
4.2 Chinese	125
4.3 Cross-linguistic comparison	126
5 Experiment design	127

128 **6 Introduction**

129 These instructions are for authors submitting
130 papers to *ACL conferences using L^AT_EX. They
131 are not self-contained. All authors must follow
132 the general instructions for *ACL proceedings,¹
133 and this document contains additional instruc-
134 tions for the L^AT_EX style files.

135 The templates include the L^AT_EX source of
136 this document (`acl_latex.tex`), the L^AT_EX
137 style file used to format it (`acl.sty`), an
138 ACL bibliography style (`acl_natbib.bst`),
139 an example bibliography (`custom.bib`), and
140 the bibliography for the ACL Anthology
141 (`anthology.bib`).

142 **7 Engines**

143 To produce a PDF file, pdfL^AT_EX is strongly
144 recommended (over original L^AT_EX plus
145 dvips+ps2pdf or dvi_{pdf}). The style file
146 `acl.sty` can also be used with luaL^AT_EX
147 and XeL^AT_EX, which are especially suitable
148 for text in non-Latin scripts. The file
149 `acl_lualatex.tex` in this repository provides
150 an example of how to use `acl.sty` with either
151 luaL^AT_EX or XeL^AT_EX.

152 **8 Preamble**

153 The first line of the file must be

154 `\documentclass[11pt]{article}`

155 To load the style file in the review version:

156 `\usepackage[review]{acl}`

157 For the final version, omit the `review` option:

158 `\usepackage{acl}`

159 To use Times Roman, put the following in
160 the preamble:

161 `\usepackage{times}`

162 (Alternatives like `txfonts` or `newtx` are also ac-
163 ceptable.)

164 Please see the L^AT_EX source of this document
165 for comments on other packages that may be
166 useful.

167 Set the title and author using `\title` and
168 `\author`. Within the author list, format mul-
169 tiple authors using `\and` and `\And` and `\AND`;
170 please see the L^AT_EX source for examples.

¹<http://acl-org.github.io/ACLPUb/formatting.html>

Command	Output
<code>{\`a}</code>	ä
<code>{\`e}</code>	ê
<code>{\`i}</code>	ì
<code>{\`I}</code>	Î
<code>{\o}</code>	ø
<code>{\'u}</code>	ú
<code>{\aa}</code>	å
Command	Output
<code>{\c c}</code>	ç
<code>{\u g}</code>	ğ
<code>{\l}</code>	ł
<code>{\~n}</code>	ñ
<code>{\H o}</code>	ő
<code>{\v r}</code>	ř
<code>{\ss}</code>	ß

Table 1: Example commands for accented characters, to be used in, *e.g.*, BibT_EX entries.

By default, the box containing the title and author names is set to the minimum of 5 cm. If you need more space, include the following in the preamble:

`\setlength\titlebox{<dim>}`

where `<dim>` is replaced with a length. Do not set this length smaller than 5 cm.

9 Document Body

9.1 Footnotes

Footnotes are inserted with the `\footnote` command.²

9.2 Tables and figures

See Table 1 for an example of a table and its caption. **Do not override the default caption sizes.**

As much as possible, fonts in figures should conform to the document fonts. See Figure 2 for an example of a figure and its caption.

Using the `graphicx` package graphics files can be included within figure environment at an appropriate point within the text. The `graphicx` package supports various optional arguments to control the appearance of the figure. You must include it explicitly in the L^AT_EX preamble (after the `\documentclass` declaration and before `\begin{document}`) using `\usepackage{graphicx}`.

²This is a footnote.



Figure 2: A figure with a caption that runs for more than one line. Example image is usually available through the `mwe` package without even mentioning it in the preamble.

9.3 Hyperlinks

Users of older versions of \LaTeX may encounter the following error during compilation:

```
\pdfendlink ended up in different
nesting level than \pdfstartlink.
```

This happens when $\text{pdf}\text{\LaTeX}$ is used and a citation splits across a page boundary. The best way to fix this is to upgrade \LaTeX to 2018-12-01 or later.

9.4 Citations

Table 2 shows the syntax supported by the style files. We encourage you to use the `natbib` styles. You can use the command `\citet` (cite in text) to get “author (year)” citations, like this citation to a paper by [Gusfield \(1997\)](#). You can use the command `\citep` (cite in parentheses) to get “(author, year)” citations ([Gusfield, 1997](#)). You can use the command `\citealp` (alternative cite without parentheses) to get “author, year” citations, which is useful for using citations within parentheses (e.g. [Gusfield, 1997](#)).

A possessive citation can be made with the command `\citeposs`. This is not a standard `natbib` command, so it is generally not compatible with other style files.

9.5 References

The \LaTeX and $\text{Bib}\text{\TeX}$ style files provided roughly follow the American Psychological Association format. If your own bib file is named `custom.bib`, then placing the following before any appendices in your \LaTeX file will generate the references section for you:

```
\bibliography{custom} 231
```

You can obtain the complete ACL Anthology as a $\text{Bib}\text{\TeX}$ file from <https://aclweb.org/anthology/anthology.bib.gz>. To include both the Anthology and your own .bib file, use the following instead of the above.

```
\bibliography{anthology,custom} 237
```

Please see Section 10 for information on preparing $\text{Bib}\text{\TeX}$ files.

9.6 Equations

An example equation is shown below:

$$A = \pi r^2 \quad (1) \quad 242$$

Labels for equation numbers, sections, subsections, figures and tables are all defined with the `\label{label}` command and cross references to them are made with the `\ref{label}` command.

This an example cross-reference to Equation 1.

9.7 Appendices

Use `\appendix` before any appendix section to switch the section numbering over to letters. See Appendix A for an example.

10 $\text{Bib}\text{\TeX}$ Files

Unicode cannot be used in $\text{Bib}\text{\TeX}$ entries, and some ways of typing special characters can disrupt $\text{Bib}\text{\TeX}$ ’s alphabetization. The recommended way of typing special characters is shown in Table 1.

Please ensure that $\text{Bib}\text{\TeX}$ records contain DOIs or URLs when possible, and for all the ACL materials that you reference. Use the `doi` field for DOIs and the `url` field for URLs. If a $\text{Bib}\text{\TeX}$ entry has a URL or DOI field, the paper title in the references section will appear as a hyperlink to the paper, using the `hyperref` \LaTeX package.

Limitations

Since December 2023, a “Limitations” section has been required for all papers submitted to ACL Rolling Review (ARR). This section should be placed at the end of the paper, before the references. The “Limitations” section (along with, optionally, a section for ethical

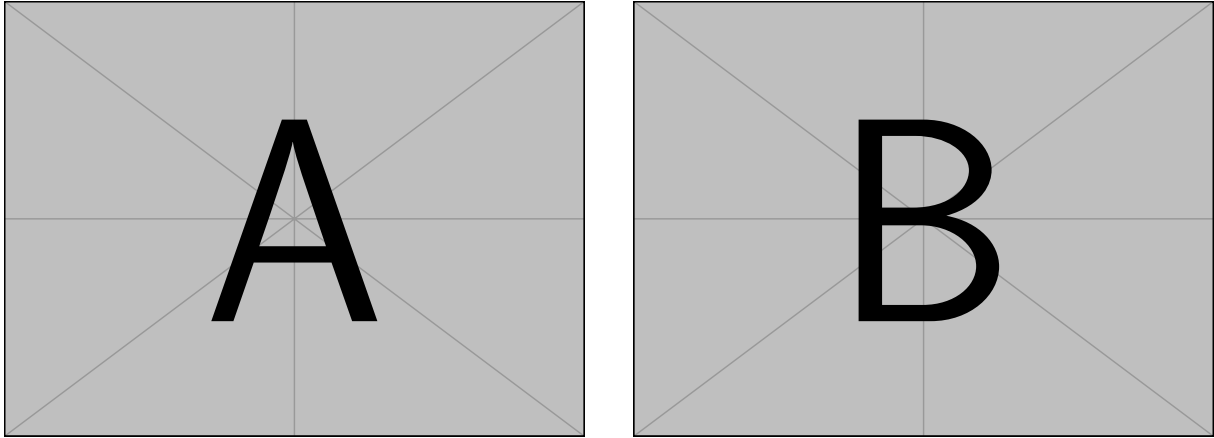


Figure 3: A minimal working example to demonstrate how to place two images side-by-side.

Output	natbib command	ACL only command
(Gusfield, 1997)	<code>\citep</code>	
Gusfield, 1997	<code>\citealp</code>	
Gusfield (1997)	<code>\citet</code>	
(1997)	<code>\citeyearpar</code>	
Gusfield’s (1997)		<code>\citeposs</code>

Table 2: Citation commands supported by the style file. The style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

considerations) may be up to one page and will not count toward the final page limit. Note that these files may be used by venues that do not rely on ARR so it is recommended to verify the requirement of a "Limitations" section and other criteria with the venue in question.

Acknowledgments

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTEX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier

ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

- Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.
- Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.
- Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.
- Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. *Yara parser: A fast and accurate dependency parser*. *Computing Research Repository*, arXiv:1503.06733. Version 2.

A Example Appendix

This is an appendix.