A Diachronic Analysis of Gender-Neutral Language on wikiHow

Katharina Suhr   **Michael Roth**
Institute for Natural Language Processing
University of Stuttgart
suhr.katharina@gmail.com, michael.roth@ims.uni-stuttgart.de

Abstract

As a large how-to website, wikiHow's mission is to *empower every person on the planet to learn how to do anything.*[1] An important part of including everyone also linguistically is the use of gender-neutral language. In this short paper, we study in how far articles from wikiHow fulfill this criterion based on manual annotation and automatic classification. In particular, we employ a classifier to analyze how the use of gender-neutral language has developed over time. Our results show that although about 75% of all articles on wikiHow were written in a gender-neutral way from the outset, revisions have a higher tendency to add gender-specific language than to change it to inclusive wording.

1 Introduction

Gender-neutral language, also known as gender-inclusive language, has its roots in the 1970s, when second-wave feminists criticized the generic use of 'he' and of gendered job titles (Hord, 2016). The demand for including women linguistically, by using gender-neutral language, has steadily increased since then. Beyond that, gender-neutral language further benefits individuals who identify outside the gender binary or when the gender of the person talked about is unknown (Hord, 2016).

The online platform wikiHow claims to be "the world's leading how-to website".[2] But is it also leading in terms of using gender-neutral language? Similar to Wikipedia, wikiHow articles can be edited publicly and all changes are stored in a revision history. A main difference is that articles are not only written by volunteers, but also by wikiHow's own experts, possibly suggesting that editing criteria also include aspects of inclusive language. In this work, we study whether this is

the case based on articles written in English, the primary language used on wikiHow.

Many articles, such as *How to Pack for a Holiday*, address the reader directly, using the gender-neutral pronoun 'you'. However, there are also articles showing that gender-neutral language is not implemented by all editors. For example, the article *How to Address a Congressman* uses the term 'congressman' in the title and the gendered phrase 'congressman and congresswoman' throughout the article text. Even though this phrase avoids the generic masculine, it is still not gender-neutral as it may not address, for instance, individuals outside the binary. A gender-neutral replacement here would have been the term 'congressperson'.

In general, different factors may contribute to the implementation of gender-neutral language in instructional texts. As a first step towards their analysis, this work seeks to answer the following questions: 1) How common are gender-neutral articles in wikiHow? 2) How did the ratio change over time? 3) Are specific users responsible for corresponding revisions?

2 Related Work

Among the first papers to include discussions of gender-neutral language for queer identities, Cao and Daumé III (2020) studied how non-binary pronouns (singular they/them and neo-pronouns) are handled by co-reference resolution systems. For this, they created two new datasets: one on "English Wikipedia about people with non-binary gender identities" and one on "articles from LGBTQ periodicals, and fan-fiction stories from Archive Of Our Own". Their results indicate that system performance significantly drops for their curated data, relative to results reported on other datasets.

Sun et al. (2021) and Vanmassenhove et al. (2021) created systems to rewrite gendered text into gender-neutral language. Both focused on using

---

[1] http://www.wikihow.com/wikiHow:Mission, accessed 6 December 2023

[2] http://www.wikihow.com/wikiHow:About-wikiHow, accessed 6 December 2023

GENDERED	Texts that uses words or phrases associated with binary gender, e.g. 'he', 'she', 'he or she', 'chairman', 'congressman', 'girlfriend/boyfriend'
GENDER-NEUTRAL	Texts that uses words and phrases that are inclusive of all genders, e.g. 'they', 'them', 'chairperson', 'partner'
NO GENDER	Texts that only show words that are not associated with any gender, e.g. 'you', 'I'

Table 1: Labels used in the annotation of the gold dataset.

they/them as neutral pronouns, as well as switching words with lexical gender to a neutral version. The words that had to be changed were defined by a static list. In contrast, Bartl and Leavy (2022) created a method that uses online dictionaries to determine the lexical gender of words. Both of these methodologies to identify words to change the gender are relevant to our work on classifying articles into gender-neutral and gendered language.

Other challenging tasks, in the landscape of gender and language studies, are Sexist Language Detection (Rodríguez-Sánchez et al., 2021, 2022) and Heteronormative Language Detection. They identify specific aspects of language that can additionally lead to bias. In the sexist language detection shared task (EXIST 2021, EXIST 2022) the goal is to identify hostile, subtle and/or benevolent sexism in English and Spanish tweets towards women (Rodríguez-Sánchez et al., 2021, 2022). In contrast, the goal of the heteronormativity language detection is to identify heteronormative assumptions in a text. Heteronormativity is a "social, political and economic regimen [where] the only acceptable and normal form to express sexual and affective desires (...) is heterosexuality" (Vásquez et al., 2022).

Instructional Text have, among other things, been used to analyzing their structure to create instructional text and answer how-to questions (Aouladomar and Saint-Dizier, 2005; Delpech and Saint-Dizier, 2008) or extract procedural knowledge (Zhang et al., 2012). An online-platform that offers a variety of instructional texts on different topics is wikiHow. WikiHow has served as a source of information for numerous research papers. For example, to detect a users' intent (Zhang et al., 2020) or to create a summarization tool (Koupaee and Wang, 2018).

In their paper, Anthonio et al. (2020) investigated how edits of users can improve texts. If they only improve the instructions' style and correctness, or if they also provide clarifications needed to follow the instructions and achieve the goal. They addressed various types of revision in their paper, such as spelling/grammar, paraphrase, information deletion, and information modification/insertion. However, neither of these types explicitly address gender or gender-neutral language. Gender-neutral language is especially important in how-to guides that are addressed to a general audience.

3 Data and Annotation

During the beginning of this work, in February 2023, wikiHow still offered the *Export pages* service, also referenced by Anthonio et al. (2020), for downloading articles and revision histories. We were able to scrape $11,074,729$ versions of a total of $256,455$ articles using this service.

We selected a small subset of these articles to first manually annotate whether gender-neutral language is used. Following our original intuition that gender-neutral language may get implemented in articles over time, we specifically searched for article versions in which the phrase 'gender' appears in the comment of a revision. From this set, 129 articles were selected.

Due to the length of the full articles, we split them into paragraphs for the annotation. One of the authors annotated each paragraph using the labels GENDERED, GENDER-NEUTRAL and NO GENDER, following the definitions provided in Table 1. A second annotator also annotated 20 articles for reproducibility and quality control. The agreement between the author and the second annotator was very high, with $\kappa = 0.912$ (Cohen, 1960).

A total of $2,247$ paragraphs were annotated, with 725 labelled GENDERED and $1,235$ labelled GENDER-NEUTRAL/NO GENDER. On the article level, we combine the labels as follows: If there is at least one paragraph with the label GENDERED, the full article is labeled GENDERED. All the remaining cases are labeled GENDER-NEUTRAL. As a result of this step, 29 annotated articles are labelled GENDER-NEUTRAL, while the remaining 100 are GENDERED.

	Precision	Recall	F$_1$
MAJORITY	0.601	0.775	0.677
PRONOUNS	0.849	0.837	0.842
STATIC LIST	**0.890**	**0.884**	**0.870**
INFERENCE	0.854	0.860	0.849
COMBINED	0.869	0.868	0.853

Table 2: Classification performance on our data. All metrics represent ***weighted averages*** across both classes.

	Precision	Recall	F$_1$
MINORITY	0.225	1.000	0.367
PRONOUNS	0.618	**0.724**	**0.667**
STATIC LIST	**0.938**	0.517	**0.667**
INFERENCE	0.789	0.517	0.652
COMBINED	0.875	0.483	0.622

Table 3: Classification performance on our data. All metrics for **GENDER-NEUTRAL** as the 'positive' class

4 Pilot Study

For classification, different rule-based and supervised variants were compared. Since gender-neutral language can broadly be defined in terms of specific features, we focus on the following rule-based classifiers:

- PRONOUNS uses regular expressions to identify gendered pronouns in an article version.

- STATIC LIST compares the content of an article to a pre-defined list of gendered words, which we collect from an online source[3] and previous work (Vanmassenhove et al., 2021).

- INFERENCE uses an online dictionary to infer the lexical gender of each noun (if any) that occurs in an article version, using code made available by (Bartl and Leavy, 2022).

- COMBINED is a combination of the previous two classifiers, labeling each paragraph as GENDERED if at least one term has a lexical gender or appears in the static word lists.

Finally, we also experimented with different supervised classifiers, but we did not observe any improvements over the rule-based classifiers.

As shown in Table 2, the classifier with the highest overall scores is the STATIC LIST classifier, with a weighted F$_1$-score of 0.87. The other three classifiers achieve comparable results to each other but perform 2–3 percentage points worse than STATIC LIST in terms of F$_1$-score.

The unbalanced setting, with 100 gendered articles out of 129 (77.5%), makes it particularly easy to identify the majority class. In Table 3, we show unweighted scores for the minority class, GENDER-NEUTRAL. As shown by the results,

---

[3] https://ielts.com.au/australia/prepare/article-grammar-101-feminine-and-masculine-words-in-english, accessed 11 December 2023

the PRONOUNS classifier achieves a higher recall, while the STATIC LIST classifier has a higher precision. Weighing precision and recall equally leads to the same GENDER-NEUTRAL F$_1$-score for both classifiers, namely a harmonic mean of 0.667. As STATIC LIST performs better for the majority class as well as in terms of weighted average scores, we use STATIC LIST in the next steps of this work.

In an error analysis, we found that one issue of STATIC LIST and other rule-based classifiers is that gendered terms can also be used as meta language, which should be classified as GENDER-NEUTRAL. For instance, some articles discuss topics related to transgender or queer issues and what terms can/cannot be used in what contexts: "An example of misgendering would be using she/her pronouns for someone who actually uses they/them, or assuming somebody with long hair is a girl." (from the article *How to Avoid Misgendering*).

5 Analysis

We apply the best-performing classifier from our pilot study, STATIC LIST, to all article revisions collected in the creation of our data (§3). The following subsections discuss three analyses to answer the questions outlined in Section 1. First, we examine the overall distribution of GENDERED and GENDER-NEUTRAL articles according to their latest version (§5.1). We then take a look at how this distribution changed over time (§5.2). Finally, we investigate the direction of revisions and check how different editors contributed to it (§5.3).

5.1 Status Quo

Given the last versions of all articles as of February 2023, our best-performing classifier labels 74% of them as GENDER-NEUTRAL. We observe a large variance regarding the use of gender-neutral language across the 19 high-level categories of wikiHow. These categories and their statistics can be
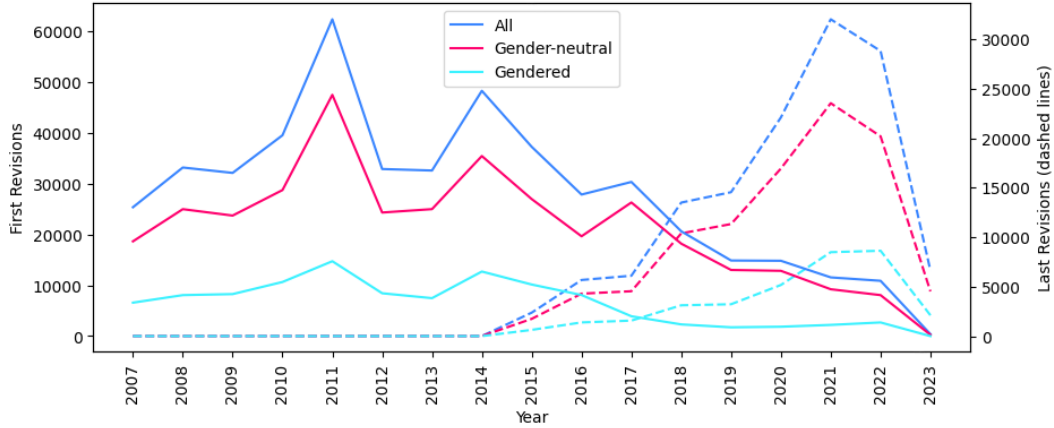
Figure 1: Overview of classifications of the first and last version of each article between 2004 and 2023.

Category	Revisions	GN
Food and Entertaining	842.390	91%
Computers and Electronics	105.4193	90%
Home and Garden	595.592	85%
Cars & Other Vehicles	249.852	80%
Hobbies and Crafts	1.169.876	73%
Sports and Fitness	427.633	69%
Travel	55.566	69%
Personal Care and Style	736.606	64%
Finance and Business	485.902	64%
Education and Communications	1.084.077	59%
Arts and Entertainment	1.123.195	59%
Holidays and Traditions	96.430	57%
Work World	73.905	57%
Health	1.157.745	55%
Pets and Animals	473.948	42%
Family Life	255.824	32%
Philosophy and Religion	132.769	30%
Youth	518.256	27%
Relationships	518.848	16%

Table 4: Percentage of classified revisions, that were classified as Gender-Neutral, separated by their categories.

found in Table 4. In most categories, the majority of articles are classified as GENDER-NEUTRAL, including for example *Computers and Electronics* (90%) and *Hobbies and Crafts* (73%). In contrast, only a minority of articles in the categories *Family Life* (32%), *Youth* (27%) and *Relationships* (16%) are GENDER-NEUTRAL.

5.2 Changes over Time

Grouping revisions together based on their article offered the opportunity to analyze the revision history of each article. As mentioned above, 74% of the last version of articles were classified as gender-neutral. But in their initial version, we found 76.4% of all articles to be classified as GENDER-NEUTRAL, which implies a decrease of 2.4 percentage points over time.

Even though there is an overall decrease in the proportion of gender-neutral articles, Figure 1 shows that there has been substantial variation over the years. In 2017, for example, the number of new GENDER-NEUTRAL articles increased while the number of new GENDERED articles decreased. In contrast, we find fewer GENDER-NEUTRAL articles last updated in 2022 in comparison to 2021, whereas the number of articles classified as GENDERED in both years stayed roughly the same.

5.3 Direction of Revisions

For each article, we compare each version's classification to the preceding one. This offers the opportunity to analyze revisions to GENDER-NEUTRAL language as well as additions of GENDERED language. In general, an article can go through multiple or no changes of label. The article *How to Put Hot Outfits Together*, for instance, saw a total of 12 changes but the article both started out as GENDER-NEUTRAL in 2007 and its last version from 2019 is still classified as GENDER-NEUTRAL.

Although GENDER-NEUTRAL versions are the majority, there are slightly more changes (51.6%) to GENDERED than revisions to GENDER-NEUTRAL. Even when examining these revisions grouped together by contributor, it is clear that most
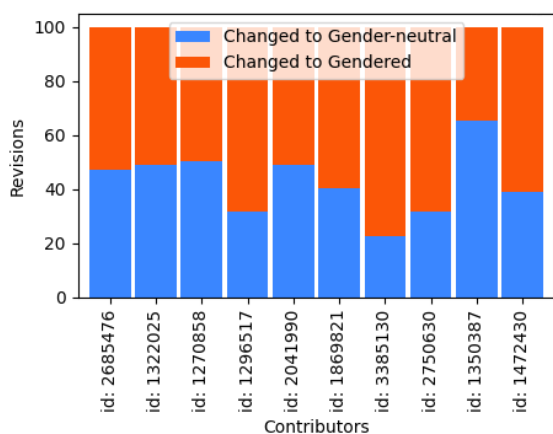
Figure 2: Ids of top-10 contributors and corresponding percentages of classified revisions. Only 1350387 performed more changes towards GENDER-NEUTRAL.

contributors are adding gender-specific language rather than revising articles to be gender-neutral. For example, Figure 2 shows that all top-8 contributors either changed more articles to GENDERED or made an equal number of edits in either direction. Among the top 10, only the second to last contributor changed substantially more articles to GENDER-NEUTRAL than to GENDERED language.

6 Conclusion

The objective of this work was to analyze the gender-neutral language of instructional texts. For this, a new dataset of revisions from wikiHow articles was created. The annotated gold dataset consists of 129 selected versions of how-to guides, 100 of which are gendered and 29 gender-neutral.

A comparison of different classifiers, mostly inspired by previous work, showed that a STATIC WORD LIST performed best on our data. A main advantage of static word lists is the option to clearly define which words are considered gendered or gender-neutral, making classifications simple and explainable. In contrast, other classifiers, such as LEXICAL INFERENCE may pick up on features associated with biological sex when detecting GENDERED language, which can lead to misidentification of binary and non-binary trans individuals.

Finally, we classified and analyzed a dataset of over 256, 000 wikiHow articles with a total of more than 11 million article versions. Our findings discussed in Section 5 suggest that, even though most articles start out as gender-neutral, there has been no concentrated effort of editors to change gendered article versions to be gender-

neutral. Nonetheless, we found several revisions in our annotation study, in which editors implemented gender-neutral language and explicitly mentioned this in the comment of the revision.

Limitations

The work presented in this paper exclusively analyzes texts written in English. Because natural and grammatical gender is encoded differently across various languages, the selected classification approach and its results are not directly applicable how-to guides written in other languages.

Furthermore, the findings in this paper are limited to one specific platform, namely wikiHow. Our results may not generalize to other platforms or to guides written for specialized topics, such as board game manuals or recipe books. Future work should address in how far the same trends can be observed outside of wikiHow.

Acknowledgements

References

Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikiHowToImprove: A resource and analyses on edits in instructional texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, page 5721–5729, Marseille, France. European Language Resources Association.

Farida Aouladomar and Patrick Saint-Dizier. 2005. Towards Generating Procedural Texts: An Exploration of their Rhetorical and Argumentative Structure. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, Aberdeen, Scotland. Association for Computational Linguistics.

Marion Bartl and Susan Leavy. 2022. Inferring Gender: A Scalable Methodology for Gender Detection with Online Lexical Databases. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, page 47–58, Dublin, Ireland. Association for Computational Linguistics.

Yang Trista Cao and Hal Daumé III. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 4568–4595. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Estelle Delpech and Patrick Saint-Dizier. 2008. Investigating the Structure of Procedural Texts for Answering How-to Questions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Levi C. R. Hord. 2016. Bucking the Linguistic Binary: Gender Neutral Language in English, Swedish, French, and German. *Western Papers in Linguistics*, 3(11).

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of EXIST 2021: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2022. Overview of EXIST 2022: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural*, 69:229–240.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, Them, Theirs: Rewriting with Gender-Neutral English.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender Neutral Alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 8940–8948. Association for Computational Linguistics.

Juan Vásquez, Gemma Bel-Enguix, Scott Thomas Andersen, and Sergio-Luis Ojeda-Trueba. 2022. HeteroCorpus: A Corpus for Heteronormative Language Detection. page 225–234, Seattle, Washington. Association for Computational Linguistics.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Intent Detection with WikiHow. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 328–333, Suzhou, China. Association for Computational Linguistics.

Ziqi Zhang, Philip Webster, Victoria Uren, Andrea Varga, and Fabio Ciravegna. 2012. Automatically Extracting Procedural Knowledge from Instructional Texts using Natural Language Processing.