# Inferring Gender: A Scalable Methodology for Gender Detection with Online Lexical Databases

**Marion Bartl**
Insight SFI Research Centre
for Data Analytics
School of ICS
University College Dublin
marion.bartl@insight-centre.org

**Susan Leavy**
Insight SFI Research Centre
for Data Analytics
School of ICS
University College Dublin
susan.leavy@ucd.ie

## Abstract

This paper presents a new method for automatic detection of gendered terms in large-scale language datasets. Currently, the evaluation of gender bias in natural language processing relies on the use of manually compiled lexicons of gendered expressions, such as pronouns and words that imply gender. However, manual compilation of lists with lexical gender can lead to static information if lists are not periodically updated and often involve value judgements by individual annotators and researchers. Moreover, terms not included in the lexicons fall out of the range of analysis. To address these issues, we devised a scalable dictionary-based method to automatically detect lexical gender that can provide a dynamic, up-to-date analysis with high coverage. Our approach reaches over 80% accuracy in determining the lexical gender of words retrieved randomly from a Wikipedia sample and when testing on a list of gendered words used in previous research.

## 1 Introduction

There is a growing body of research on gender bias embedded in trained language models as well as on allocational and representational harms caused by the deployment of these models. There have moreover been increasing calls for early and thorough data description and curation in order to gain insights into how, for instance, gender stereotyping or quality of service bias is propagated from data into a language model. What both of these strands of research have in common is their reliance on pre-defined lexicons of terms related to gender.

In English, gendered words most commonly include pronouns (*he*, *she*, *they*, etc.), and also words that carry lexical gender, such as *boyfriend*, *policewoman*, or *prince*. Previous works on gender bias in language technologies often use manually compiled lists of words carrying lexical gender to, for example, mitigate gender stereotyping through data augmentation (Lu et al., 2020), assess

trans-exclusionary bias in co-reference annotations (Cao and Daumé III, 2020) or evaluate gender inequalities in Wikipedia article titles (Falenska and Çetinoğlu, 2021). However, curated lists are limited in their coverage of terms that contain lexical gender and can become outdated if not maintained.

To address this issue, we present a scalable algorithmic method to determine lexical gender by querying a word's dictionary definitions for a small subset of definitively gendered words. Our method allows for high-coverage, instantaneous detection of words carrying lexical gender, which eliminates the need to manually compile and maintain static lists of gendered words. This not only facilitates the extension of previous work on gender bias in NLP, but can also be used for a more detailed analysis on the representation of gender in large-scale language datasets used to train large language models like BERT (Devlin et al., 2019) or GPT-2 (Radford et al., 2019).

By combining the gender labels obtained from Merriam Webster Online (Merriam-Webster, 2022), WordNet® (Princeton University, 2010) and Dictionary.com (Dictionary.com, LLC, 2022), our method reaches an accuracy of 84% in determining the lexical gender of words in a random sample of 1,000 Wikipedia articles and 87% accuracy on a list of words carrying lexical gender adapted from previous research. The code for the algorithm, evaluation methods and datasets are available[1].

In the following section we first outline the conceptions of linguistic gender used in this research and secondly present an overview of research on gender in language technology that relies on curated lists of gendered words. Thirdly, we discuss prior approaches to algorithmic gender inference. Section 3 gives a detailed overview of the algorithm and Section 4 introduces the datasets used to assess our gender detection algorithm. We present

---

[1] https://github.com/marionbartl/lexical-gender

quantitative and qualitative results in Section 5 and discuss limitations as well as avenues for future development.

## 2 Background

When dealing with the category of gender in language technology, it is important to make a distinction between the social category of gender and gender in a linguistic sense. While social gender relates to the complex property, performance and experience of one's own and others' gender within society (Ackerman, 2019), linguistic gender describes the expression of gender within grammar and language. In English, linguistic gender mainly encompasses ways to express gender as female, male or gender-indefinite (Fuertes-Olivera, 2007). Social gender, as an extra-linguistic category, includes a more fluid view of gender aside from male and female categories. This includes transgender, genderqueer and other non-binary experiences and expressions of gender (Darwin, 2017). As Bucholtz (1999) and Cao and Daumé III (2020) point out, there is no "one-to-one" mapping between social and linguistic gender. However, the two are influenced by each other: on one hand, expressions of gender in language are subject to changing norms in society (Fuertes-Olivera, 2007), on the other hand, the way gender is represented in language influences the conception of gender within society (Butler, 1990). Thus, being able to evaluate gendered expressions in language provides insights into societal conceptualisations of gender.

Since this research explicitly focuses on lexical gender in English, which is a linguistic category, we give an overview of linguistic gender in English in Section 2.1. Section 2.2 explores the role lexical gender information plays in different areas of research on gender bias in NLP, which simultaneously present possible areas of application for our method of lexical gender inference. Section 2.3 discusses two prior algorithmic systems for lexical gender inference in English.

### 2.1 Linguistic gender in English

The taxonomy of linguistic gender in this work builds upon the approach developed by Cao and Daumé III (2020) and incorporates work by Corbett (1991), Hellinger and Bussmann (2003) and Fuertes-Olivera (2007).

Within linguistic gender, Cao and Daumé III (2020) differentiate between grammatical, refer-

ential, and lexical gender. **Grammatical gender** refers to the distinction of noun classes based on agreement between nouns and their dependants. English, as a natural or notional gender language (McConnell-Ginet, 2013), does not have grammatical gender, but it has referential and lexical gender. **Referential gender** is used to refer to the social gender of a specified extra-linguistic entity. Thus, it "relates linguistic expressions to extra-linguistic reality, typically identifying referents as 'female', 'male', or 'gender-indefinite.' " (Cao and Daumé III, 2020). In English, pronouns fall under the category of referential gender. **Lexical gender**, which this work focuses on, is non-referential but a semantic property of a given linguistic unit, which can be either masculine, feminine[2] or gender-indefinite/gender-neutral. Ackerman (2019) calls these words "definitionally gendered". Words that carry lexical gender can require semantic agreement in related forms, such as, for instance, using the pronoun *his* in connection with the word *stuntman* in the sentence 'Every *stuntman* needs to rehearse *his* stunts.' (Fuertes-Olivera, 2007). In English, lexical gender is usually not morphologically marked. Exceptions to this rule include e.g. the suffixes *-man* to denote masculine gender, such as in *policeman*, or *-ess* to denote feminine gender, such as in *waitress*. It should moreover be noted that lexical gender is exclusively a linguistic property. However, words containing lexical gender can be used to express referential gender if a concrete referent is specified (Cao and Daumé III, 2020).

### 2.2 Lexical gender in gender bias research

The evaluation and mitigation of gender biases in language datasets and models relies on referential expressions of gender, such as pronouns, but also words that carry lexical gender. These pieces of research vary in application, as well as the number of gendered expressions considered, which varies from two to around 120 words. Most works assess binary differences between male and female gender. However, an emergent strand of NLP research also focuses on non-binary gender expressions (Cao and Daumé III, 2020) and creating gender-neutral datasets and systems (Vanmassenhove et al., 2021). The following considers example use-cases of lexicons of lexically gendered words. These simultaneously represent a variety of applications for our

---

[2]We use the terms *masculine* and *feminine* instead of *male* and *female* here in order to underline the purely linguistic, i.e. semantic, property of lexical gender

lexical gender detection algorithm.

**Dataset evaluation**   The most straightforward form of using gendered words is to assess the distribution of gendered words in a corpus. Zhao et al. (2019) counted *he/she* pronouns in the One Billion Word Benchmark (Chelba et al., 2013) to show male skew in the training data for the ELMo language model (Peters et al., 2018), which is the primary focus of their analysis. This analysis addressed calls for better data evaluation (Bender et al., 2021; Rogers, 2021) prior to or alongside model bias analyses.

**Retrieval for analysis**   Limited-scope lists of words that carry lexical gender were used by Caliskan et al. (2017) to retrieve Word2Vec embeddings (Mikolov et al., 2013) and perform the Word Embedding Association Test (WEAT). This test measured stereotyping by calculating implicit associations between eight male/female word pairs and words related to maths or science and arts. Guo and Caliskan (2021) used an adapted version of the WEAT, the CEAT, to asses intersectional biases in contextualized word embeddings (ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), OpenAI GPT (Radford et al., 2019; Brown et al., 2020)). Another use-case in which gendered words were used for retrieval is research by Falenska and Çetinoğlu (2021), who assessed gender bias in Wikipedia articles. As a first step, they filtered the article titles for a limited number of words that carry lexical gender.

**Creation of synthetic evaluation data**   In sentence-based analyses of gender-bias, lists of words with lexical gender can also be used to fill placeholders in sentence templates and thus create synthetic sentences with different gendered entities. For example, Kiritchenko and Mohammad (2018) created the Equity Evaluation Corpus (EEC) to analyse gender stereotyping in sentiment analysis systems. The EEC inspired the creation of the Bias Evaluation Corpus with Professions (BEC-Pro), which was used to analyse associations between gendered entities and professions in BERT (Bartl et al., 2020). Similarly, Sheng et al. (2019) used the word pair *the man/the woman* as fillers within sentence-start prompts for open-ended natural language generation (NLG) and the subsequent analysis of gender biases in the generated sentences.

In a rare instance of research on non-binary representations of gender in NLP, Cao and Daumé III

(2020) used gendered lists of words to find and hide lexical gender in the GAP dataset (Webster et al., 2018). The dataset created in this way was used to measure gender- and trans-exclusionary biases in coreference resolution performed by both humans and machine-learning models.

**Data manipulation**   Extensive lists of gendered words were used in the context of Counterfactual Data Augmentation (CDA), which replaces words with masculine lexical gender with their feminine variants and vice versa in a corpus. This is done in order to create training or fine-tuning data for gender bias mitigation. For instance, Lu et al. (2020) 'hand-picked' gender pairs to swap in CDA and Maudslay et al. (2019) added first names to the list of words to be swapped.

Another kind of data manipulation, this time aiming for neutral gender, was performed by Vanmassenhove et al. (2021). They used lists of unnecessarily gendered job titles (e.g. *mailman/mailwoman*) and feminine forms (e.g. *actress*), as well as generic uses of the suffix *-man* (such as in *freshman*) in the extended version of their *Neutral Rewriter*, which re-writes explicit mentions of gender into their gender-neutral variants (*mail carrier*, *actor*, *first-year student*).

## 2.3   Lexical gender inference

Previous approaches to automatic lexical gender inference used unsupervised and semi-supervised learning, drawing on the presence of gendered pronouns in the context of a given noun (Bergsma and Lin, 2006; Bergsma et al., 2009). While Bergsma and Lin (2006) created a large dataset of probabilistic noun gender labels, Bergsma et al. (2009) used these as basis for creating training examples for a statistical model that uses context and morphological features to infer lexical gender.

One major point of criticism here lies in the probabilistic determination of noun gender, which has the risk of mislabelling lexically neutral nouns, such as professions, as being gendered due to contextual distributions that are representative of stereotypes or the number of men and women holding the profession instead of the linguistic category of lexical gender. For example, since there are more female than male nurses (Bureau of Labor Statistics (BLS), 2022) and thus most nurses are referred to with female pronouns in text, the algorithm might infer that the term *nurse* has female lexical gender, when in fact it is neutral.
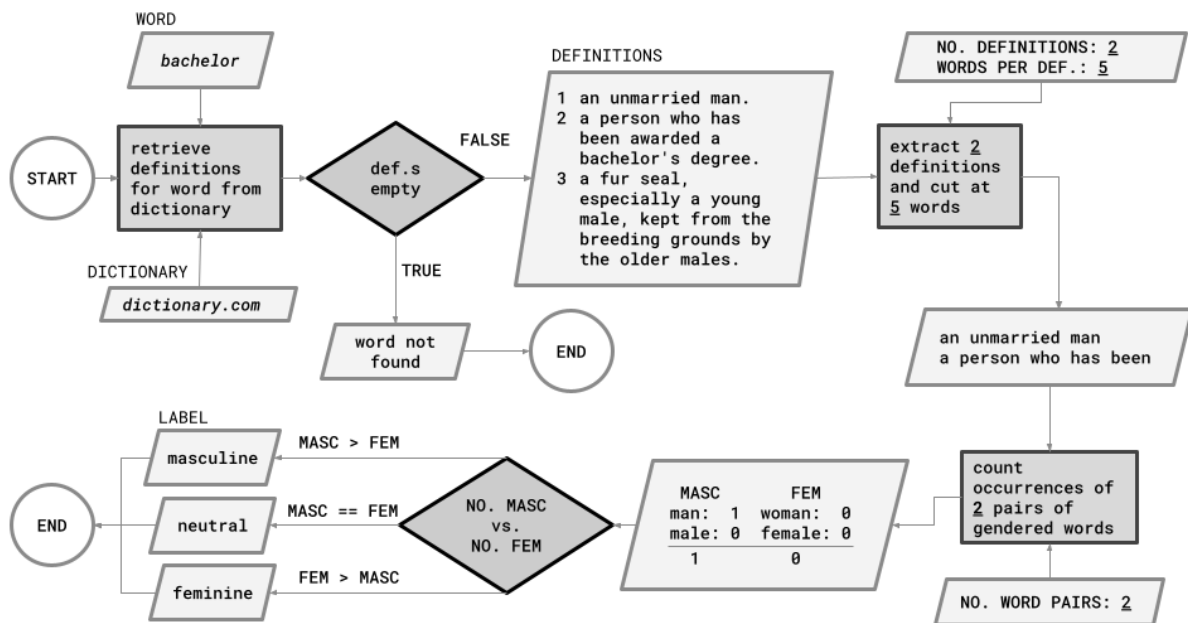
Figure 1: Simplified exemplary flowchart of gender detection algorithm

## 3 Method: Automatic Detection of Lexical Gender

The main goal of this work is to produce a dynamic, high coverage, scalable method to determine the lexical gender of a target word in order to replace previously used manually compiled lexicons. For this purpose, we leveraged the fact that the definition of a lexically gendered word includes words from a small set of definitively gendered words that carry the same lexical gender. In the following, we describe the main algorithm setup, additional parameters and heuristics, as well as the method to combine lexical gender labels from different databases. A schematic, exemplary overview of the algorithm is presented in Figure 1.

### 3.1 Algorithm construction

The method we outline utilises the increasing availability of machine-readable dictionaries, such as Merriam Webster Online, Dictionary.com, and the lexical database WordNet, in order to identify gendered terms. Examples (1) and (2) illustrate how lexical gender is captured within Merriam-Webster's (2022) definitions of *nun* and *monk*:

(1) *nun*: a woman belonging to a religious order

(2) *monk*: a man who is a member of a religious order and lives in a monastery

Both definitions mention the lexical gender of the referent through a gendered word, in this case *man* and *woman*. Initial analyses showed that gendered words are more likely to occur at the beginning of a definition and definitions often used the words *female/male* or *woman/man* to specify lexical gender. In identifying gendered terms, we thus considered the presence and amount of up to eight definitively gendered words, such as *male/female*, *man/woman* etc., in the target word's definitions to draw inferences about its lexical gender.

For retrieval of the definitions, we accessed WordNet through the Natural Language Toolkit (NLTK, Bird et al., 2009) and Merriam Webster Online as well as Dictionary.com through HTTP requests.

Once the definitions for a given target word were retrieved, the process of obtaining lexical gender was the same for either dictionary. We determined whether a word has masculine, feminine, or neutral lexical gender by counting occurrences of a number of word pairs which have clearly defined feminine or masculine lexical gender, which are displayed in Table 1. If the combined definition texts contain more masculine than feminine terms, the word was labelled with masculine lexical gender, and vice versa. If the same number of masculine and feminine words was found within a set of definitions, which includes the case in which none of the pre-

| $w$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **feminine** | woman | female | wife | daughter | mother | girl | sister | aunt |
| **masculine** | man | male | husband | son | father | boy | brother | uncle |

Table 1: Words carrying explicit lexical gender; $w$ = number of pairs used for experiments

defined gendered terms can be found, the word was labelled with neutral lexical gender. We additionally obtained a combined label through a majority vote over the individual dictionaries' gender labels. In cases in which words could not be found in one dictionary and querying each of the other dictionaries returned different labels, a neutral gender label was assigned.

## 3.2 Parameters

Three variable parameters were used to limit the number of definitions and word tokens queried, as well as the number of definitively gendered words to use for the query. In order to determine the best combination of values for our parameters, we performed a grid search using our gold standard data (see Section 4.1) and combined labels to test performance.

**Number of definitions** $d$   We limited the number of definitions, because definitions that occur early on have a higher likelihood of describing a more general sense of the word, while later definitions relate to very specific word senses. Therefore, we retrieved only the first $d$ definitions that the dictionary lists for the word. During grid search, we tested integer values in the range $d = [2..10]$, and the best value was determined to be $d = 4$.

**Number of tokens** $t$   We also experimented with limiting the number of tokens within a given definition to see whether definitively gendered terms were more likely to be mentioned earlier in a given definition. The definitions were tokenized using NLTK (Bird et al., 2009). We took the first $t$ tokens of each definition. Regarding the number of tokens in a definition, we tested the algorithm with $t = \{5, 10, 15, 20, 25, 30, 35\}$ in our experiments and found $t = 20$ to produce optimal results.

**Number of gendered word pairs** $w$   The word pairs used during experiments are listed in Table 1. The first two word pairs, *woman/man* and *female/male*, as well as the pair *girl/boy*, are most commonly used to describe the gender of a person or animal, while the rest of the words describes

gendered family relations. The latter were chosen in order to account for cases in which the lexical gender of a person is described in relation to another person by using family terms. This is for example the case in the definition of *baroness* in Merriam Webster: "the wife or widow of a baron" (Merriam-Webster, 2022). The grid search was performed for integer values in the range $w = [2..8]$ and best performance was obtained for $w = 5$ word pairs. Moreover, if a target word was included in the definitively gendered pairs or their plural forms, it was automatically classified with the respective lexical gender.

## 3.3 Morphological Heuristics

Aside from the lexical database method described above, we additionally applied heuristics relating to suffix-morphology and punctuation. Morphological heuristics were applied before querying the dictionaries, while the punctuation-related heuristic was applied when a word could not be found in a dictionary.

The first heuristic was applied in order to handle gender-neutral definitions of words that carry gender-explicit markers, such as the word *businessman*, which carries the masculine suffix *-man*. Its definition in WordNet (Princeton University, 2010) is shown in (3).

(3) *businessman*: a person engaged in commercial or industrial business (especially an owner or executive)

Even though *businessman* contains a masculine suffix, its definition is generic, most likely due to the fact that *businessman* was once used for business people of all genders. However, since feminine or neutral equivalents (*business woman, business person*) are widely used nowadays, the word *businessman* has become gender specific and defining it generically represents an outdated, male-as-norm viewpoint (Fuertes-Olivera, 2007).

We thus classified words containing the suffixes *-man* and *-boy* or *-woman* and *-girl* into masculine and feminine lexical gender, respectively. Regular

| | gold (N=134) | Wiki1000-sample (N=515) | | | Wiki1000 dataset (N=12,643) | | |
|---|---|---|---|---|---|---|---|
| POS | NN | NN | NNS | all | NN | NNS | all |
| **masc** | 53 | 82 | 43 | 125 | 100 | 46 | 146 |
| **fem** | 53 | 51 | 29 | 80 | 60 | 28 | 88 |
| **neut** | 28 | 212 | 98 | 310 | 7,679 | 3,880 | 11,559 |
| **not found** | - | - | - | - | 618 | 232 | 850 |
| **all** | 134 | 345 | 170 | 515 | 8,457 | 4,186 | 12,643 |

Table 2: Composition of evaluation corpora for lexical gender detection algorithm.
**Note**: for *Wiki1000 full*, combined predicted labels were used, because no gold labels exist for this dataset

expressions were used to ensure that feminine or neutral words ending in *-man* such as *woman* or *human*, as well as words that have the suffix *-woman*, were not classified as masculine.

Another heuristic was applied in order to account for spellings that differ in punctuation, e.g. *grandfather* vs. *grand-father*. We check for and subsequently remove punctuation within a word if it cannot be found within a dictionary. This also applies to the cases in which non-detection is caused by a whitespace character.

## 4 Data

We used two test datasets to evaluate and run the algorithm. The first dataset, which we call *gold standard* hereafter, contains nouns that have a clear lexical gender and were mainly sourced from previous research on gender bias. The second dataset contains 1,000 randomly sampled Wikipedia articles, which we used to extract gendered nouns. The following describes both datasets in detail.

### 4.1 Gold Standard

In order to gain insights into the performance of the dictionary-based algorithm for lexical gender retrieval, we compiled a list of words that have a nearly unambiguous lexical gender, which acts as the *gold standard*. The gold standard list was developed based on a lexical gender list by Cao and Daumé III (2020) with the addition of more words retrieved from online lists for learners of English[3][4][5]. Nouns retrieved from prior research and online sources were subsequently filtered for explicitness of lexical gender. For example, the

pair *actor/actress* would not be considered since the word *actor* is nowadays used for both male and female referents. We moreover added neutral gender replacements for word pairs for which such an alternative exists. An example would be the triplet *headmaster*-MASC, *headmistress*-FEM, *headteacher*-NEUT. The final list is comprised of 53 masculine, 53 feminine, and 28 neutral words (see Table 4 in the Appendix).

### 4.2 Wikipedia Sample

This research aims at providing a flexible, scalable, and high-coverage method for lexical gender detection. Therefore we additionally tested the approach on more naturalistic data, namely a random sample of 1,000 articles from English Wikipedia obtained through the *wikipedia* python library[6]. We will abbreviate this sample corpus as *Wiki1000* hereafter.

The articles were then cleaned and tokenized into sentences using NLTK (Bird et al., 2009) and subsequently processed with SpaCy to obtain part-of-speech (POS) tags for each word. All singular and plural nouns (POS-tags: NN, NNS) were then extracted and analysed for lexical gender. Nouns that contained special characters due to cleaning and tokenization errors were dropped. This method provided us with 12,643 nouns, as illustrated under Wiki1000 in Table 2.

In order to test the performance of the algorithm, the instances of the Wiki1000 dataset needed true labels. A corpus size of 12,643 instances, however, was beyond the scope of this research to manually label. In fact, it represents the kind of corpus size that we aim to label automatically. We therefore filtered Wiki1000 for nouns that were labelled as either masculine or feminine by Merriam Webster Online, Dictionary.com, or WordNet. Like this, we

---

[3]www.vocabularypage.com/2017/03/gender-specific-nouns.html
[4]7esl.com/gender-of-nouns/
[5]learnhatkey.com/what-is-gender-in-english-grammar/

[6]https://pypi.org/project/wikipedia/

|          | gold standard (N=134) | | | | Wiki1000-sample (N=515) | | | |
|----------|------|------|------|----------|------|------|------|----------|
| measure  | **P** | **R** | **F1** | **Acc** | **P** | **R** | **F1** | **Acc** |
| **WordNet** | 0.91 | 0.83 | 0.85 | **0.83** | 0.73 | 0.63 | 0.63 | **0.63** |
| **Merriam Webster** | 0.89 | 0.77 | 0.8 | **0.77** | 0.83 | 0.82 | 0.82 | **0.82** |
| **Dictionary.com** | 0.93 | 0.87 | 0.89 | **0.87** | 0.76 | 0.61 | 0.59 | **0.61** |
| **Combined** | 0.92 | 0.87 | 0.89 | **0.87** | 0.85 | 0.84 | 0.84 | **0.84** |

Table 3: Quantitative results for lexical gender detection of gold standard and Wiki1000-sample

specifically target gendered nouns and obtain a corpus similar to the gold standard corpus, but sourced from naturally occurring text. The resulting corpus of 515 nouns, which we call *Wiki1000-sample*, was subsequently labelled for 'true' lexical gender by members of the research team (Fleiss's $\kappa \approx 0.87$). The labels used for evaluation were determined by majority vote. The specifications of the Wiki1000-sample dataset can be found in Table 2.

In line with previous research on gender bias in Wikipedia (Wagner et al., 2015; Falenska and Çetinoğlu, 2021), which found an over-representation of male entities in the encyclopedia, Table 2 shows that there are approximately 1.5 times as many mentions of distinct entities with masculine lexical gender in our 1,000-article Wikipedia sample than there of entities with feminine lexical gender.

## 5 Results and Discussion

### 5.1 Quantitative analysis

An overview of algorithm performance on the gold standard dataset and the reduced Wiki1000 sample can be found in Table 3. We report the weighted average of precision, recall, and F1-measure due to unbalanced classes in our test data.

As seen in Table 3, our best performing approach on both the gold dataset (87% accuracy) as well as the sample of Wiki1000 (84% accuracy) was combining labels from all three sources by majority vote. Keeping in mind that the Wiki1000 sample is approximately three times the size of the gold standard, the relative consistency in performance here indicates robustness for our approach. It should also be noted that only querying Dictionary.com reached the same performance on the gold standard dataset (87% accuracy) while on the Wiki1000 sample, using only Merriam Webster reached a comparable accuracy score to the combined model (82%).

Table 3 moreover shows that on the gold standard dataset, which was used to fine-tune our parameter values using grid search, our method reached an accuracy of 77% or higher in each experiment configuration. Using the same parameter values for experiments on the Wiki1000 sample, only the combined approach as well as using only Merriam Webster reaches an accuracy of >77%. When using only WordNet or Dictionary.com, the performance drops from 84% to 63% and 61% accuracy, respectively. This shows that parameter configurations can be adapted to specific dictionaries and dataset sizes.

Figure 2 shows confusion matrices for the combined approach on both the gold standard dataset (2a) and the Wiki1000-sample (2b). Figure 2a shows that on the gold standard, the combined classifier mislabelled four feminine and 11 masculine instances as neutral, but did not mislabel any of the neutral instances as either masculine or feminine. In contrast, both these classification mistakes can be found on the Wiki1000 sample (Figure 2b). Here, the algorithm classifies more lexically neutral words as gendered than vice versa.

Cases in which lexically neutral words are classified as gendered include words that are traditionally related to specific genders, such as *bikini* or *soprano*, as well as *patriarchy* or *testes*. It is likely that dictionary definitions reflect this traditional gender association, leading to misclassification. Conversely, classifications of gendered words as neutral can e.g. be caused by definitions that do not mention gender, either because of presumed knowledge (*pope*) or because a lexically specific word was formerly used for all genders (*landlord*). Another reason for gendered-as-neutral misclassification can be the definition of one gendered term by using another, which 'cancel each other out'. For example, WordNet defines *widow* as "a woman whose husband is dead especially one who has not

remarried" (Princeton University, 2010).

Another issue, which only occurred when testing on the gold standard dataset, concerns words that could not be found. The first is *single person*, which we chose as gender-neutral alternative for *bachelor/spinster*. The fact that it was not found could be due to the term *single person* being more of a composite phrase than a joined expression. Moreover, single people are often described using the adjective *single* in a predicative way, such as in the sentence 'He is single.', instead of 'He is a single person.' The other word that could not be found is *child-in-law*, which is the gender-neutral variant of *son/daughter-in-law*. Here, the issue could be frequency of use, since *child-in-law* is less established than its gender-specific variants.

## 5.2 Qualitative analysis

The following section discusses some classification errors in more detail. We focus on errors that occur due to gender-exclusive definitions in the lexical databases caused by historically close associations of words to a single gender.

In our first example, an outdated definition in WordNet (Princeton University, 2010) causes the misclassification of the word *crew*, a neutral term, as masculine. We show the first and fourth definitions in Example (4) in order to illustrate how the masculine label was obtained.

(4) *crew*

  1. the men and women who man a vehicle (ship, aircraft, etc.)
  4. the team of men manning a racing shell

In the first definition, the words *men and women* are used to define the crew of any vehicle while in the fourth definition, which describes the crew of a racing shell (a type of rowing boat), only the word *men* is used. This leads to a masculine lexical gender label, since the definitions taken together contain more masculine than feminine words. However, the fourth definition could have been worded like the first, or used the word *people*, since racing shells can be crewed by people of any gender.

A similar classification error occurred for the words *soprano*, *menopause* and *nurse*, which were all classified as feminine by the combined model, even though they have neutral lexical gender. These terms are all closely associated with female social gender due to anatomical and hormonal differences between sexes (*soprano* and *menopause*), historical biases of women performing care-work, as well as current gender distributions in certain professions (*nurse*; Bureau of Labor Statistics (BLS), 2022). While using gender-exclusive wording to define lexically neutral terms could inform readers of a word's traditional relation to social gender, it can also reproduce gender stereotypes and exclude those who do not identify as female but still sing in soprano voice or work as a nurse. Moreover, using feminine words in the definition of words like *menopause* can be seen as a form of trans-exclusionary bias, since people assigned female at birth, whose body can cease to menstruate, might not identify as female.

## 5.3 Limitations and Future Developments

We have selected dictionaries to obtain the lexical gender of a word, because they represent a relatively objective resource that is expected to list neutral and non-stereotypical definitions of words. However, as shown in Section 5.2, dictionaries are after all a human-curated resource and as such still carry human biases and outdated definitions, which in turn lead to biased or outdated results.

We would moreover like to point out that we are explicitly working with English, which does not mark gender grammatically. In languages that mark grammatical gender, our method would most likely be obsolete, because here gender can e.g. be inferred from formal features such as morphology or agreement for most nouns (Corbett, 1991). What is more, English, as a lingua franca and the language most focused on by the NLP community (Bender et al., 2021), has a plethora of high-quality and high-coverage resources available. Since our method is reliant on lexical resources, adapting the method to low-resource languages could prove challenging. However, while more complex lexical resources like WordNet might not yet exist for some languages, it is likely that online dictionaries do exist. Therefore, we still believe that our method can be adapted to other notional gender languages (McConnell-Ginet, 2013).

Another limitation of the present work concerns word sense disambiguation, since the presence of lexical gender depends on the word's sense in context. As an example, the word *colt*, can either mean a young male horse or a brand of pistol. In the sense of a male horse, the lexical gender of *colt* is clearly masculine while in the sense of the pistol, it
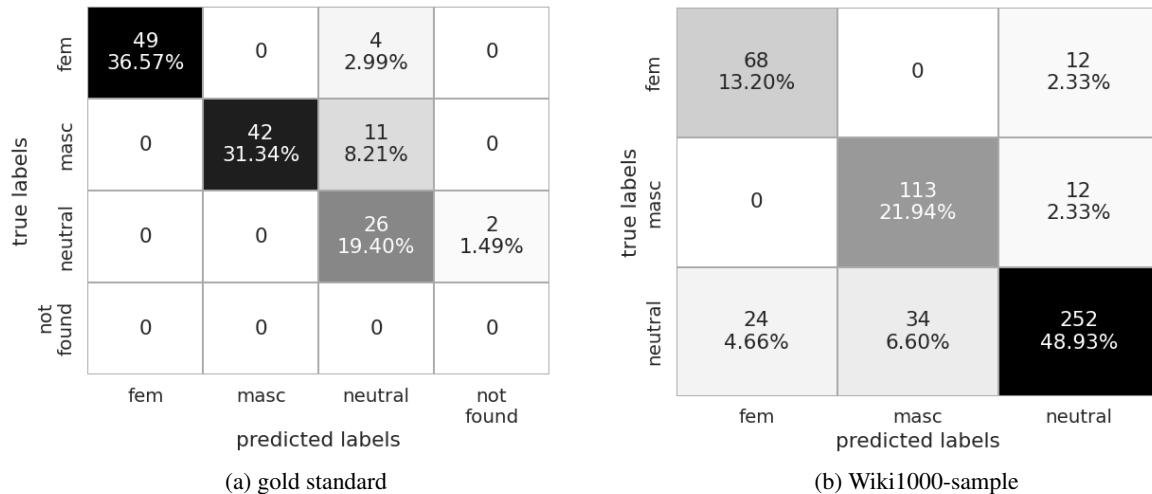
Figure 2: Confusion matrices for combined labels
words that were not found in (a): *single person*, *child-in-law*

is neutral. Differences in the lexical gender of word senses can also be caused by semantic shifts, such as for the word *master*, which traditionally refers to a man who is in control of e.g. servants or a household. However, in an academic context its meaning has shifted and now refers to an academic degree, or more broadly to a person of undefined gender who has reached a high level of skill in a given discipline. Therefore, future work will integrate word sense disambiguation within the algorithm.

## 6 Conclusion

We have presented a method to automatically determine the lexical gender of a given word by querying its dictionary definitions. The performance of the algorithm on a gold standard dataset of gendered nouns based on related literature, as well as a set of nouns sampled from 1,000 randomly selected Wikipedia articles, reached up to 87% accuracy. Previous research on gender bias in NLP used manually compiled lists of gendered words for data evaluation, retrieval, manipulation, and the synthetic creation of data. In contrast, our method is scalable and has a high, dynamic coverage, which gives it a variety of applications within past and future research on gender bias in NLP. These include e.g. the assessment of gender representations in large-scale corpora, the retrieval of gendered words for which gender-neutral replacements need to be found, as well as determining whether male-centric language such as epicene *he* is used in coreference resolution clusters.

## References

Lauren M Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa*.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. Conference Proceedings.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics*, pages 33–40.

Shane Bergsma, Dekang Lin, and Randy Goebel. 2009. Glen, glenda or glendale: Unsupervised and semi-supervised learning of english noun gender. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 120–128.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Mary Bucholtz. 1999. Gender. *Journal of Linguistic Anthropology*, 9(1/2):80–83.

Bureau of Labor Statistics (BLS). 2022. Labor Force Statistics from the Current Population Survey.

Judith Butler. 1990. *Gender trouble: feminism and the subversion of identity*. Book, Whole. Routledge, London; New York.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. Publisher: American Association for the Advancement of Science.

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*, pages 4568–4595, Online. Association for Computational Linguistics.

Ciprian Chelba, Tomás Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, and Phillipp Koehn. 2013. One billion word benchmark for measuring progress in statistical language modeling. *CoRR*, abs/1312.3005.

Greville G. Corbett. 1991. *Gender*. Cambridge University Press.

Helana Darwin. 2017. Doing gender beyond the binary: A virtual ethnography. *Symbolic Interaction*, 40(3):317–334.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dictionary.com, LLC. 2022. Dictionary.com.

Agnieszka Falenska and Özlem Çetinoğlu. 2021. Assessing Gender Bias in Wikipedia: Inequalities in Article Titles. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 75–85, Online. Association for Computational Linguistics.

Pedro A. Fuertes-Olivera. 2007. A corpus-based view of lexical gender in written Business English. *English for Specific Purposes*, 26(2):219–234.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133.

Marlis Hellinger and Hadumod Bussmann. 2003. *Gender across languages: the linguistic representation of women and men*, volume 11. J. Benjamins, Amsterdam; Philadelphia.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275.

Sally McConnell-Ginet. 2013. Gender and its relation to sex: The myth of 'natural' gender. In Greville G. Corbett, editor, *The Expression of Gender*, pages 3–38. De Gruyter Mouton.

Merriam-Webster. 2022. Merriam-webster.com.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Princeton University. 2010. About WordNet.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Anna Rogers. 2021. Changing the World by Changing the Data. *arXiv:2105.13947 [cs]*. ArXiv: 2105.13947.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

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*, pages 8940–8948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media*.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

# Appendix

| category | masculine | feminine | neutral |
|---|---|---|---|
| family | brother | sister | sibling |
| | dad | mum | |
| | dad | mom | |
| | daddy | mummy | |
| | daddy | mommy | |
| | father | mother | parent |
| | father-in-law | mother-in-law | parent-in-law |
| | fiance | fiancee | betrothed |
| | grandfather | grandmother | grandparent |
| | grandson | granddaughter | grandchild |
| | husband | wife | spouse |
| | nephew | niece | |
| | son | daughter | child |
| | son-in-law | daughter-in-law | child-in-law |
| | step-father | step-mother | step-parent |
| | stepfather | stepmother | stepparent |
| | uncle | aunt | |
| | widower | widow | |
| misc | bachelor | spinster | single person |
| | boy | girl | child |
| | boyfriend | girlfriend | partner |
| | gentleman | lady | |
| | groom | bride | |
| | lad | lass | |
| | male | female | |
| | man | woman | person |
| | manservant | maidservant | servant |
| | steward | stewardess | attendant |
| | swain | nymph | spirit |
| | wizard | witch | |
| occupation | businessman | businesswoman | business person |
| | chairman | chairwoman | chairperson |
| | fireman | firewoman | fire fighter |
| | headmaster | headmistress | head teacher |
| | landlord | landlady | renter |
| | milkman | milkmaid | |
| | policeman | policewoman | police officer |
| | salesman | saleswoman | salesperson |
| | waiter | waitress | server |
| religion | friar | nun | |
| | monk | nun | |
| title | Mr. | Mrs. | Mx. |
| | baron | baroness | |
| | count | countess | |
| | czar | czarina | |
| | duke | duchess | |
| | earl | countess | |
| | emperor | empress | ruler |
| | king | queen | |
| | prince | princess | |
| | signor | signora | |
| | sir | madam | |
| | viscount | viscountess | |

Table 4: Masculine, feminine and neutral nouns of the gold standard dataset