TÉCNICO LISBOA

TRANSdutor: A Rewriting Approach for Gender Inclusivity in Portuguese

Leonor Silva Pereira de Sousa Veloso

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisors: Prof. Maria Luísa Torres Ribeiro Marques da Silva Coheur
Prof. Rui Orlando Magalhães Ribeiro

Examination Committee

Chairperson: Prof. João António Madeiras Pereira
Supervisor: Prof. Maria Luísa Torres Ribeiro Marques da Silva Coheur
Member of the Committee: Prof. Maria Inês Camarate de Campos Lynce Faria

November 2023

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

Gostaria de agradecer aos meus orientadores, Luísa e Rui, por me darem a liberdade de perseguir um tema que me é tão querido. Nenhuma palavra deste trabalho existiria sem a vossa incansável ajuda, apoio, e incentivo.

Aos meus pais, que me deram todos os recursos e fizeram todos os sacrifícios possíveis para eu chegar até aqui. Estarei eternamente grata pela vossa paciência, amor, e constante apoio numa fase em que não tive tanto tempo e atenção para vos dar como mereciam. À minha avó, que desde pequena me inspirou.

Ao Fortuna e à Teresa, que aos meus olhos sempre cá estiveram e sempre cá estarão, e me estão a ver a terminar ainda outra fase da vida. Espero que vejam muitas mais.

À Lúcia e à Mariana, que deram todos estes passos ao meu lado, e sem as quais não tenho dúvida que teria caído pelo caminho. Ao Aylton que, caso eu efetivamente caísse, não me deixaria bater com a cabeça no chão. Ao Biscoito, que me deu a confiança para ir avante com este tema e para muito mais. À Dani, por me ter acolhido no meu primeiro ano como uma *queer* ainda verde e me ter ensinado todos os termos do glossário desta tese. Ao Regouga e ao Rocha, por todos os dias me fazerem rir e ouvirem todos os meus lamentos e histórias do quotidiano. Ao Baltasar, que sempre me inspirou a ser melhor, de todas as maneiras possíveis. Ao Diogo, à Lara, e à Matilde, por, no inóspito clima austríaco, me terem ouvido divagar sobre género gramatical sem se queixarem.

Aos incontáveis amigos que fiz ao longo deste percurso — em particular à Lídia, à Teresa, ao Albino, ao Afonso, à Constança, ao Patrício, à Sofia, ao Ulisses — por constantemente me aconselharem, ouvirem, e por terem partilhado comigo este troço de vida que me marcará para sempre.

Aos anotadores e outros que contribuíram para este trabalho e o tornaram uma realidade, muito obrigada.

Abstract

In recent years, there has been a notable rise in research interest regarding the integration of gender-inclusive and gender-neutral language in Natural Language Processing models. A specific area of focus that has gained practical and academically significant interest is *gender-neutral rewriting*, which involves converting binary-gendered text to its gender-neutral counterpart. However, current approaches to gender-neutral rewriting for gendered languages tend to rely on large datasets, which may not be an option for languages with fewer resources, such as Portuguese. In this thesis, we present a rule-based and a neural-based tool for gender-neutral rewriting for Portuguese, a heavily gendered Romance language whose morphology creates different challenges from the ones tackled by other gender-neutral rewriters. Our neural approach relies on fine-tuning large multilingual Machine Translation models on examples generated by the rule-based model. We evaluate both models on texts from different sources and contexts. Results show that both rule-based and neural approaches reach a similar level of performance, although the rule-based model performs marginally better in most types of text. We provide the first Portuguese dataset explicitly containing gender-neutral language and neopronouns, as well as a manually annotated golden collection of 500 sentences that allows for the evaluation of future work.

Keywords

Gender-neutral Rewriting, Bias, Gender Bias, Machine Translation, Natural Language Processing

Resumo

A integração de linguagem género-neutra e inclusiva em modelos de Processamento de Linguagem Natural é um tópico de interesse na literatura atual. Um tópico específico que tem ganho tracção e vindo a ser de particular interesse prático e teórico é a *reescrita de linguagem género-neutra* (*gender-neutral rewriting*, em Inglês). Esta tarefa consiste em converter linguagem que apenas contém pronomes masculinos ou femininos — os pronomes binários — em linguagem género-neutra. As abordagens atuais para esta tarefa tendem a depender de um grande volume de dados, o que pode não ser uma abordagem viável para linguagens que possuem menos recursos, tal como é o caso do Português. Nesta tese, apresentamos dois modelos que abordam a tarefa de reescrita de linguagem género-neutra: um modelo baseado em regras e um modelo neuronal. A nossa abordagem neuronal consiste em afinar grandes modelos multilíngues de Tradução Automática, utilizando como dados de treino exemplos gerados pelo modelo baseado em regras. Avaliamos ambos os modelos em frases de diferentes fontes e contextos. As contribuições desta tese consistem na primeira coleção de dados em Português que contém explicitamente linguagem género-neutra e neopronomes, bem como uma coleção dourada de 500 frases manualmente anotadas que permitem a avaliação deste trabalho e de possível trabalho futuro.

Palavras Chave

Reescrita de Linguagem Género-neutra, Bias, Bias de Género, Tradução Automática, Processamento de Língua Natural

Contents

List of Figures

List of Tables

Acronyms

RBM Rule-Based Model

NM Neural Model

OWN-PT OpenWordnet-PT

NLI Natural Language Inference

CR Coreference Resolution

NLP Natural Language Processing

POS Part-Of-Speech

NMT Neural Machine Translation

MT Machine Translation

LGBTQ+ Lesbian, Gay, Bisexual, Transgender, Queer, and more

LLM Large Language Model

WER Word Error Rate

CER Character Error Rate

Glossary

butch

Notably or deliberately masculine in appearance or manner (Merriam-Webster, 2023).

cisgender

Of, relating to, or being a person whose gender identity corresponds with the sex the person had or was identified as having at birth (Merriam-Webster, 2023). Antonym of transgender.

drag

Entertainment in which performers caricature or challenge gender stereotypes (Merriam-Webster, 2023).

gender

The behavioral, cultural, or psychological traits typically associated with one sex (Merriam-Webster, 2023).

gender identity

A person's internal sense of being male, female, some combination of male and female, or neither male nor female (Merriam-Webster, 2023).

gender-neutral

Not referring to either sex but only to people in general (e.g. *gender-neutral language*) (Merriam-Webster, 2023).

genderqueer

Of, relating to, or being a person whose gender identity cannot be categorized as solely male or female (Merriam-Webster, 2023).

neopronouns

Coined pronouns (Hekanaho, 2020). English examples include *ze, xe*. Portuguese examples include *ile, elu*.

non-binary

Relating to or being a person who identifies with or expresses a gender identity that is neither entirely male nor entirely female (Merriam-Webster, 2023).

sex

Either of the two major forms of individuals that occur in many species and that are distinguished respectively as female or male especially on the basis of their reproductive organs and structures Merriam-Webster (2023).

transgender

Of, relating to, or being a person whose gender identity differs from the sex the person had or was identified as having at birth (Merriam-Webster, 2023). Antonym of cisgender.

1

Introduction

Contents

The relationship between language and gender, as well as its effects in societal gender dynamics, has been examined and documented as early as the 70s (Gal, 1989, 1978). In more recent years, there has been a push towards the usage of gender-neutral (or *gender-fair*) language. How this gender-fairness is achieved, is, however, highly dependent on the specifics of each language. During the course of this work, we will concern ourselves with the case for Portuguese.

Like most Romance languages, Portuguese is characterized by a binary grammatical gender system in which nouns belong to one of two classes: masculine or feminine. During the last decade, with the advent of social media, there has been an increase in the visibility and usage of new sets of gender-neutral pronouns (Pinheiro, 2020) (sometimes referred to as *neo-pronomes* in Portuguese). These neopronouns are preferred by many individuals to refer to themselves, but can also be used to refer to a mixed-gender group of people avoiding the default masculine plural (for example, *um grupo de alunes* instead of *um grupo de alunos*). The usage of neopronouns is becoming increasingly unavoidable and preferred by many in order to write and speak in a gender-inclusive Portuguese (Miranda, 2020).

Current Portuguese Natural Language Processing (NLP) models tend to ignore this shift in language and the phenomena of neopronouns. The lack of gender inclusion in NLP datasets is, of course, a multilingual problem (Zhou et al., 2019). Nevertheless, there are efforts being made towards the processing of gender-neutral language, perhaps most notably in English (Vanmassenhove et al., 2021; Sun et al., 2021). As such, we set out to create a model that allows for the processing of Portuguese gender-neutral language in the same way that is currently being achieved for other languages.

1.1 Motivation

In recent years there has been clear societal push towards the usage of gender-inclusive and gender-neutral language. Examples include the addition of the singular English pronoun *they* to the Merriam-Webster dictionary[1] and the addition of the gender-neutral Swedish pronoun *hen* to the Swedish Academy's SAOL[2].

A 2020 study focused on the Spanish language (Slemp, 2020) found that 90% of its non-binary participants struggled with "expressing or describing their gender identity in Spanish" (compared to 3% of participants who identified as either a man or a woman). Furthermore, "only 36% of participants stated that they never had difficulty describing someone else's gender identity in Spanish". Most of the participants that used gender inclusive language (including neopronouns) claimed they had begun its use only two to five years before the study took place. These results are of special interest to us, as they show the rapidly growing need of strategies for gender inclusion in Romance languages.

[1] https://www.merriam-webster.com/
[2] https://svenska.se/saol/

Having these societal needs in mind, the motives for our work largely lie on the concepts of gender *affirmation* and *visibility*.

Gender *affirmation* is defined as "the process by which individuals are affirmed in their gender identity through social interactions (Sevelius, 2013)". The usage of the correct name and pronouns to refer to someone is included in the concept of gender affirmation and is a key component in the mental well-being of many, especially transgender individuals. It has been shown that the incorrect use of pronouns (commonly referred to as *misgendering*) may even lead HIV-positive transgender women to avoid proper healthcare (Sevelius et al., 2020). There is also growing awareness of how biases and gender dynamics present in society impact the way we construct technology, and how that technology can directly and negatively impact individuals and historically marginalized communities. A recent study sheds some light on misgendering and encoded gender bias in algorithmic systems: most of the participants said they "considered misgendering by automatic gender recognition worse than being misgendered by another human being" (Hamidi et al., 2018). Two of the participants suggested "giving people autonomy over the way they are gendered by technology".

Another concern that motivates our work is how the usage of the default masculine plural can restrict the *visibility* of other genders and invoke a *male bias* (the assumption that a person of undefined gender is a man) *Male bias*, or *masculine bias*, is a sociolinguist phenomenon whose research dates as far back as the 1980s (Sniezek and Jazwinski, 1986), and its implications extend to NLP. Research regarding the Spanish language (similar in terms of grammatical gender to Portuguese) shows that strategies for gender asymmetry alleviation reduce the cognitive male bias of participants (Kaufmann and Bohner, 2014). One of the strategies used by the quoted study was the gender neutral form *-x* (*lxs españolxs*) which is also used in Portuguese (see Section 2.3). As such, we believe that the exploration of a Portuguese neutral grammatical gender, as well as other strategies for gender-inclusive writing, are relevant for the development of gender-unbiased NLP models.

1.2 Problem

As we have established, there is a need to make sure that current-day NLP models accompany these shifts towards gender inclusion in language. Unfortunately, we face a lack of Portuguese datasets that use gender-neutral language in order to train these models. This is an especially relevant problem for the tasks of Machine Translation (MT) and Coreference Resolution (CR).

Regarding MT, the need for gender inclusive models arises most notably when we translate from non-gendered to gendered languages (and vice-versa), and maintaining gender-neutrality is required.

CR is described by Jurafsky and Martin (2021) as "the task of determining whether two mentions *corefer*, by which we mean they refer to the same entity in the discourse model (the same discourse

entity)". As such, CR usually entails the assumption of the gender of human entities in a text. If, for instance, a human entity in an English text uses *they/them* pronouns, the CR model must be able to correctly process and identify those pronouns as such.

We must also face the fact that how exactly a "neutral" grammatical gender in Portuguese should be implemented, in a morphological sense, is a constant debate, and there are several systems in use by Portuguese-speaking communities today. Creating a model that identifies and processes correctly all forms of gender-neutral pronouns (and how they change the gender of associated nouns and adjectives) in usage today would be a very extensive work. In Section 2.3 we go more in depth about this topic and present a few of the most commonly found systems in literature and in communities.

1.3 Objective & Contributions

Our objective is to develop models that allow a user to create a more gender-inclusive version of a desired Portuguese text. The three main contributions of our work consist of:

- As far as we concern, the first Portuguese parallel datasets explicitly containing gender-neutral language and neopronouns, made publicly available[3], as well as a manually curated test set of 500 sentences. These datasets are comprised of sentences belonging to five different text categories: literary texts, journalistic texts, dialogues, social media posts and comments, and simpler sentences. This allows for the evaluation of Portuguese gender-neutral rewriters in different contexts. We hope this contribution will increase visibility of gender-neutral language and neopronouns in the landscape of Portuguese NLP datasets, as well as allow for future research regarding these topics;

- A rule-based gender-neutral rewriter based on handcrafted rules, for which we provide open access[4];

- A neural gender-neutral rewriter[5] developed via fine-tuning a large multilingual machine translation model. This method requires relatively smaller sized datasets (when compared to training a model from scratch), and thus allows for the development of gender-rewriters for lower-resource languages.

We establish a baseline for the gender-neutral rewriting task for the Portuguese language. By publicly releasing our models and datasets, we hope to encourage further research in this specific area.

[3] https://github.com/leonorv/pt-gn-datasets
[4] https://github.com/leonorv/pt-gender-neutralizer
[5] https://huggingface.co/leonorv/pt-neural-gender-neutralizer

1.4 Document Outline

Chapter 2 focuses on the theoretical, social, and linguistic background of our work. In Chapter 3, we review studies that fall into the scope of our work. Chapter 4 describes how we have processed and compiled our automatically curated and manually curated datasets. These serve as training data for our neural model and test set, respectively. Chapter 5 details the architecture of our rule-based and neural models. In Chapter 6, we present our experimental setup, metrics, and evaluation results, along with a detailed error analysis. Finally, Chapter 7 summarizes our work and its limitations, as well as possible avenues for future work.

1.5 Publication

A portion of this thesis, namely the material covered in Chapters 4, 5, and 6, is featured in a paper accepted to the Findings of Empirical Methods in Natural Language Processing (EMNLP) 2023.

2

Background

Contents

This chapter goes in depth about the theoretical concepts of grammatical and "natural" gender and explores linguistic strategies to achieve *gender-fairness* in language. We present a proposal for a Portuguese neutral grammatical gender, based in recent literature and work done by the Portuguese-speaking Lesbian, Gay, Bisexual, Transgender, Queer, and more (LGBTQ+) community. The section closes with a few ethical considerations and concerns that will follow us throughout our work.

2.1 Grammatical Gender and "Natural" Gender

Understanding the linguistic concept of grammatical gender is crucial for this project, since it has implications both in human cognition and natural language processing (Corbett et al., 1991). Hockett (1967) provides us with an elegant definition: "Genders are classes of nouns reflected in the behaviour of associated words". A language that is considered to be *gendered* possesses two or more grammatical genders, and will require some level of agreement between the grammatical gender of the noun and the items that are related to it. Depending on language, those items may be articles, verbs, adjectives, or others. A noun is said to *belong* to a gender, while an adjective that is related to it is said to be *inflected* for gender (Hockett, 1967). For example, in the Portuguese sentence "A cadeira está partida" (The chair is broken), the feminine noun "cadeira" (chair) agrees with the preceding feminine article "A", and the verb termination -a in the verb "partida" (broken).

Grammatical gender systems greatly differ between languages. Portuguese, like most romance languages, has two genders (masculine and feminine), while German has three (masculine, feminine, and neutral). In most, the categories correspond at least partly to the distinction of sex, but this is not always the case (Corbett et al., 1991). It is also common for languages to distinguish genders based on whether the nouns refer to animate or inanimate objects. For example, Czech has masculine, feminine, and neutral genders, but masculine subdivides into animate and inanimate. Dyirbal, an Australian indigenous language, has four grammatical genders (Lakoff, 2008):

I: for denoting males and animals;

II: for women, water, fire, and fighting;

III: for non-flesh food;

IV: for everything not the previous categories.

Languages that are considered to be non-gendered may still apply a grammatical gender system when referring to people. These languages are sometimes referred in literature as being "natural-gender languages" (Auxland, 2020). In English, a non-gendered language, we use the pronoun *he* to refer to people that identify with the male gender (or to anyone who prefers it), (sometimes) to refer to male pets, or to refer to inanimate objects in the case of personification. Therefore, the gendered and non-gendered language divide is not strict (Konishi, 1993). In the interest of simplicity, over the course of this work we

will still use the terms *gendered* and *non-gendered* to refer to languages.

Regarding grammatical gender in the Portuguese language, nouns can belong to one of two classes: variable (*variáveis*) or invariable (*invariáveis*). Variable nouns usually share the same root, but their masculine and feminine forms differ (e.g. *aluno/aluna*). Invariable nouns have only one form, regardless of gender. However, invariable nouns can be of fixed gender (*género fixo*) or variable gender (*género variável*). Fixed gender nouns have only one form and one gender - and, as such, terms related to it (such as definite articles) must agree with the gender of the noun. Variable gender nouns also possess the same form for the masculine and feminine forms, but the gender of terms related to it (such as definite articles) can vary depending on the gender of the person we are referring to. This phenomenon is depicted in Figure 2.1.

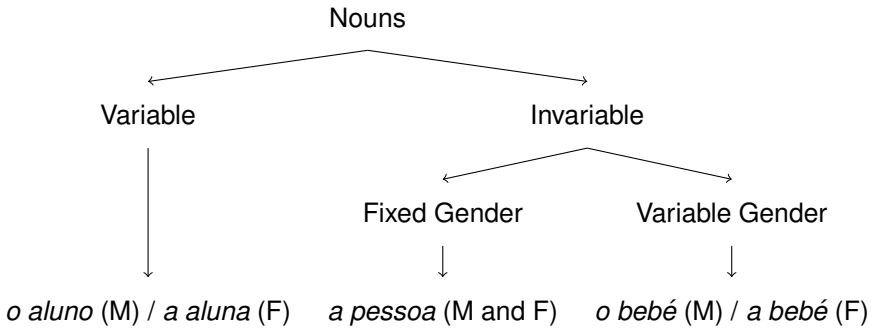

Figure 2.1: Noun classes in the Portuguese language, regarding grammatical gender. The examples for each class contain the masculine (M) and the feminine (F) form of each noun, preceded by the respective definite article.

Over the course of this work, we refer to the proposals for a Portuguese neutral gender and invariable nouns as "gender-neutral", both for simplicity and alignment with previous work done for other languages in this area of research. However, we would like to note that the Portuguese term *género-neutro*, which we use as a direct translation of the English term "gender-neutral", is used in field of linguistics to refer specifically to certain pronouns, such as *alguém*[1] or *ninguém*[2]. Although the term is often reserved for these types of pronouns, we will use "gender-neutral" to refer to terms which hold the same form regardless of the gender of the referent. As such, we will refer to terms such as *pessoa* ou *bebé* as "gender-neutral", even though we might lose some linguistic correctness in the process.

Whether grammatical gender and its assignment to nouns carries any meaning or relationship to the concept of "natural gender", meaning gender as an *attribute* or *characteristic* of people (commonly referred to as gender identity), is an ongoing discussion in the field of linguistics (Konishi, 1993). While examining why and how nouns are attributed to a certain grammatical gender falls outside the scope of our work, examining its effects on cognition and on societal dynamics might be of interest to us. It has

[1] English: "someone"
[2] English: "nobody"

been shown that grammatical gender might affect people's mental representations of objects, creating an association between gendered objects and the stereotypes related to humans of the same gender (Phillips and Boroditsky, 2003).

However, that is not the only connection between grammatical gender and natural gender. The pronouns a person uses to refer to themselves tend to be a reflection of their gender identity, or are at least influenced by the way they wish for their gender to be interpreted by society. For example, for transgender historian Stryker (2017), "One's gender identity could perhaps best be described as how one feels about being referred to by a particular pronoun". An individual might choose to use several pronouns to refer to themselves, and these might be dependent on context or on who they are speaking to. For famous historian and activist Tyroler (2006), "pronouns are always placed within context. I am female-bodied, I am a butch lesbian, a transgender lesbian – referring to me as "she/her" is appropriate, particularly in a non-trans setting in which referring to me as "he" would appear to resolve the social contradiction between my birth sex and gender expression and render my transgender expression invisible. I like the gender neutral pronoun "ze/hir" because it makes it impossible to hold on to gender/sex/sexuality assumptions about a person you're about to meet or you've just met.".

We also see grammatical gender being used as a tool for political and artistic purposes. For example, Hokenson (1988) notes the rejection of female pronouns (*she/her* in English or *ela/dela* in Portuguese) by lesbian writers across history, whether as a pseudonym or as a rejection of the binary categories of sex in language. Another example is the way in which drag performers might use different pronouns while in character and in real-life (a drag queen might use *she/her* on stage, but *he/him* in all other contexts). Of course, this depends on the individual's personal choice (Rogers, 2018).

2.2 Towards a Gender-Fair Language

Attitudes towards gender inclusivity in language are rapidly changing. Gustafsson Sendén et al. (2021) argue that, until recent years, work towards gender inclusivity in language has been focused on making women more "salient in comparison with men, or by actively avoiding androcentric language". The increasing societal awareness of non-binary identities and genders beyond the traditional masculine/feminine binary has resulted in the introduction of different strategies for gender inclusivity in language.

Sczesny et al. (2016) have identified two different strategies to make language "gender-fair": *neutralization* and *feminization*. Neutralization includes the replacement of gendered terms, whether by using a neutral word with a similar meaning (e.g., *alunos* becomes *estudantes*) or a gender-neutral form (e.g., *alunos* becomes *alunes*). Feminization consists of replacing the default masculine with a masculine-feminine pair (*alunos* becomes *alunos e alunas*), and it is often used in Portuguese and other binary-gendered languages to signal gender inclusivity. It has been shown to be a valid strategy to mini-

mize 'male bias' (the assumption that a person of undefined gender is a man) by Stahlberg et al. (2001). In one of the conducted experiments, participants were asked to name the first three famous people who came to mind, given a category. If the question was phrased using a masculine-feminine pair (Politikerinnen[feminine] und Politiker[masculine]), the answers tended to include more female names than if the question used only a masculine generic.

The coining of neopronouns and their usage as a neutralization strategy has also been shown to be successful. Gustafsson Sendén et al. (2021) studied the changes in the Swedish population's attitudes toward the neutral neopronoun *hen* from 2015 to 2018. In 2015, 61.5% of the participants claimed to never have used *hen*. In 2018, this number drops to 47.5%. However, under 10% of the participants claim to use *hen* with a high frequency ("a few times a week" or "daily"). The authors considered the increase in acceptance and usage of *hen* "noteworthy" for a time period of 3 years.

Hord (2016) published a cross-lingual study on gender-neutral language for English, Swedish, French, and German. The participants spoke any of those languages and identified as transgender, non-binary, or genderqueer. Two out of the six Swedish-speaking participants reported using *hen*, and none of them reported using any other Swedish neopronoun. 66% of the English-speaking participants used "some form of gender neutral pronoun" to refer to themselves. 34% of the English-speaking participants used the singular *they* (as a gender-neutral pronoun). English neopronouns were used at 1% or 2% each, and are considered by the author to "not being used in high concentrations despite the proliferation of them on the internet, their use in writing, and the attention they receive in the media". Two of the six bilingual French/English-speaking participants reported using neutral pronouns in English. One reported using the French neopronoun *iel*, and the other "avoided a choice by using *mon*". The German/English bilingual speakers who used neutral pronouns in English reported not using neutral neopronouns in German, but instead avoiding it in speech.

The same study provides some insight into the communities that are in fact using gender-neutral language. 44% of the English-speaking participants said that "gender neutral language was trans-specific". One confessed: "When I was using gender-neutral pronouns in English, it was almost impossible to get anyone who wasn't in the queer community to use 'they' for me consistently." The French and German participants said they thought gender-neutral language to be specific to trans communities.

Particularly in countries where the dominant language is gendered, gender-neutral language and neopronouns are still associated with queerness, and especially with trans communities. However, cases such as the increased usage of the Swedish *hen*, motivated by measures like its introduction in the SAOL dictionary, might indicate that in time gender-neutral language can reach a wider acceptance in mainstream society.

2.3 Proposals for a Neutral Grammatical Gender Within Portuguese

2.3.1 Overview

Most of the literature we found on proposals for a Portuguese neutral grammatical gender comes from practical guides, both by Brazilian and European Portuguese authors. These guides tend to be based on informal studies and observations of the neutral language used by queer communities.

The earliest source we were able to find is the 2014 Berlucci and Zanella's "Manifesto ILE", which suggests the implementation of the *ile/dile* pronouns. However, the authors of this guide do not provide any rules or recommendations on the usage of this pronoun and agreement with nouns and adjectives.

Caê (2020) presents the *Elu*, *Ile*, *Ilu*, and *El* neopronoun systems. They differ in terms of the pronoun in use, but the system for agreement with nouns and adjectives is similar, using an -e termination ("filho[3]" becomes "filhe"). The author also provides some general tips for rephrasing sentences so as to omit gender. For example,"Ela caiu[4]" becomes "Aquela pessoa caiu[5]".

The in-house work from Santos and Marques recommends the usage of the neopronoun *éle/déle* and the termination system -e (as a contrast to the masculine termination -o and the feminine termination -a). An example sentence formed with the rules from this guide is: "O professor deu as boas-vindas a todes es alunes.[6]". The authors also recommend the usage of already existent neutral forms (e.g., "monarca[7]" instead of "rei/rainha[8]").

The only other European Portuguese guide we found sets rules for a system using the pronoun *elu/delu* (Valente, 2020). The authors provide general grammatical rules and example sentences, such as "Aquelu menine é minhe filhe" (that kid is my child).

Most of the systems we observed converge on how the agreement of the neutral pronoun with nouns and adjectives should be done. For example, the neutral termination tends to be -e (*filho* becomes *filhe*), since -e is a vowel that provides contrast in speech with -o and -a. Understanding the variety of systems in use today might prove to be complex. To provide the reader with a general overview of how these systems work, we present an example sentence using different neopronouns in Table 2.1.

In Appendix A.3 we define the rules for a gender-neutral grammar that we will use during the development of our models and experiments. It is primarily based on the *Elu* system, but is heavily influenced by all the previously cited guides.

[3] English: "son"
[4] English: "She fell"
[5] English: "That person fell"
[6] English: "The teacher welcomed all the students."
[7] English: "monarch"
[8] English: "king/queen"
[9] English: "He ate his pizza."

Personal Pronoun	"Ele comeu a pizza dele.[9]"
elu	"Elu comeu a pizza delu."
ile	"Ile comeu a pizza dile."
ilu	"Ilu comeu a pizza dilu."
éle	"Éle comeu a pizza déle."
el	"El comeu a pizza del."

Table 2.1: Usage of different Portuguese neopronouns.

2.3.2 General Expectations

So far, we have seen specific systems and guides on neutral language usage, but the work of Auxland (2020) defines general expectations for a Portuguese third neutral grammatical gender:

- It should be distinct from the grammatical masculine or feminine, at least in cases where there is a grammatical gender distinction.
- It should conform, as much as possible, to the existing grammar of the Portuguese language, as a way to facilitate assimilation and acceptance into mainstream society.
- It should be easily understood by those outside queer and activist communities.
- It should "mirror existing vocabulary and linguistic practice".
- It should "function in terms of use as a singular, specific personal noun, alongside functioning as a more generalised group noun".
- It should "function in both written and spoken contexts".

Whether any of the systems we have presented fulfills all of these requirements can be subjective. For example, there is concern that an -e termination might be mistaken in speech for an -o termination (*filhe*/*filho*), but for some a system that highly differs from the existing Portuguese grammar might be too difficult to assimilate.

2.3.3 Rejection of -x and -@ Terminations

The characters -x and -@ have also been adopted as gender-neutrality markers (eg: *todxs* or *tod@s*, but there is concern with the legibility and usability of said markers. Since reading softwares cannot pronounce words that use -x or -@, their usage has been categorized as ableist (Berlucci and Zanella; Santos and Marques, 2021; Valente, 2020). Beyond that, -x and -@ are only writing markers, being

impossible to pronounce and therefore useless in speech. For these reasons, these terminations tend to be rejected by the community.

During the course of this project, we have subscribed to the rejection of -x and -@ and focused on the implementation of rules based on -e and -u terminations.

2.4 Ethical Considerations

First and foremost, we would like to acknowledge that most of the literature that was analysed for this work comes from a European/North American context, in which Portuguese culture is usually integrated. As such, it is worth noting that the terms and concepts related to gender used in this paper (such as *transgender* or *non-binary*) may not have a direct "translation" to other languages and cultures where the divide between sexual orientation and gender is not as defined. This is noted by Hord (2016) in *Bucking the Linguistic Binary*, where they provide an example of that disconnect in the Spanish language with the essay "Transliteration" (Fernández, 2010). For a language with very different language processing challenges from Portuguese, we would like to recommend *Queer Japanese: Gender and Sexual Identities through Linguistic Practices* (Abe, 2010), where the author discusses the case for Japanese. With the present work we do not propose to create a gender-neutral language system or NLP model for cultures and associated languages that do not share the concepts related to queerness and language with the Portuguese culture and language.

Secondly, we must point out that gender-neutral forms of gendered languages and neopronouns are an ever-changing phenomenon, and that we do not claim that any gender-neutral grammar system, gender-neutral termination, or neopronoun is in any way superior to others. In fact, our model should accommodate the addition of new systems. In a similar way, although we argue for their relevancy, we do not claim that any of these systems *should* be used in either writing or speech.

3

Related Work

Contents

In this chapter, we present an overview of existing gender rewriter systems, both English-focused and for other languages. We briefly cover the topic of gender bias in NLP systems, focusing on the aspects that are most relevant for our work. Additionally, we present tools and models that may be relevant both for our own work and the general task of gender-neutral rewriting for Portuguese.

3.1 Rewriter Systems for Gender Inclusivity

We find that our proposed system falls into the category of what we will call *rewriter* or *gender-neutral rewriter* systems. These rewriters take as input some form of gendered text and output a gender-neutral version of that same text. What is considered to be *gender-neutral* greatly varies between studies, and between languages. These models may be *monolingual*, where their function is to rewrite from one language to the same language (in the same fashion that we propose) or *cross-lingual*, where their purpose is to create gender-ambiguous translations when necessary. It may be worth noting that most of these models are very recent (the oldest one that includes the concept of gender-neutral pronouns being from 2021). As such, at the time of writing we were only able to find six of them, developed for a total of four languages. However, we expect that similar models continue to be developed in the following years.

3.1.1 Gender-fair rewriting for English

For the English language, perhaps the most unavoidable mention is the work of Vanmassenhove et al. (2021), who propose two models for rewriting English text with gender-inclusive and gender-neutral alternatives (singular *they*): a rule-based rewriter (RBR) and a neural rewriter (NMT). The RBR uses the Stanza Part-Of-Speech (POS) tagger and dependency parser to map binary forms to the corresponding gender-neutral alternatives ("She has her book" becomes "They have their book") and changes some gendered expressions to more inclusive ones ("chairman" becomes "chairperson"). The NMT uses a Transformer model (Vaswani et al., 2017) trained with data processed by the RBR, and was "able to generalize over the rule-based generated data, outperforming it with error rates below 0.18% (0.0% (WB+), 0.18% (OpenSubtitles) and 0.02% (Reddit)".

Also in 2021, Sun et al. published a very similar proposal for an English gender-neutral rewriter using *they*. They created both a rule-based and a neutral model, using a Transformer architecture. The neural model is trained using data processed by the rewriting algorithm. Their data is augmented by "converting a masculine sentence to a feminine sentence or vice versa and keeping the same gender-neutral translation". As for results, "the algorithm and the model achieve over 99 BLEU and less than 1% word error rate". Contrary to the work described immediately before, the algorithm was found to perform "marginally better" than the neural model. Vanmassenhove et al. speculated that this was due

to their own rule-based model having a better performance, “leading to better source (gendered)–target (neutral) training data for the NMT model”. However, according to the study itself, “nearly half of the model’s mistakes are due to rare tokens like whitespaces, emojis, and symbols”.

3.1.2 Gender-fair rewriting for Other Languages

Similar rewriters for gendered languages have been proposed. Diesner-Mayer and Seidel (2022) created a system for German. Their work consisted of a rule-based model that detects the generic masculine plural and suggests either a masculine/feminine pair or the ”gender star” (a German typographic style for gender inclusivity). The detection and correction is done in three steps, the first being the detection of masculine nouns and pronouns. The second step consists of the following checks:

- A personal designation check: all nouns that do not refer to people are discarded. The system used here is interesting and very specific to German: “a noun does not refer to a person, if a feminine declination of the noun does not exist”.
- Exclusion of forms that are already written in a fashion that the study considers gender-neutral.
- Exclusion of proper nouns and respective pronouns.

The third step consists of the application of suggestions as explained above. As for the system evaluation, “about 88 % of the occurrences were identified correctly. Grammatically well-formed suggestions were generated for about 94% of the correctly identified occurrences”, which provides an optimistic view of the performance of the rule-based models for gender-neutral rewriting. However, we must factor in that this model does not account for neopronouns. Beyond that, this model suffers from a limitation for our purposes: we wish to not only rewrite usages of the generic masculine plural, but rewrite personal pronouns according to the users’ pronoun preference.

The earliest rewriter system that we were able to find is the “Gendercheck Editor” tool developed by Carl et al. (2004). Its rule-based system checks for gender discriminatory formulations in German texts. The system uses a “marking and filtering strategy”, where possible discriminatory expressions (such as the usage of the default masculine) are marked and then pass through a selection of filters to determine if these expressions exist in a gendered context. For example, if a default masculine common name is preceded by a family name, then it is filtered out, as it refers to a single person and is not considered a default masculine. The suggestions that the system outputs to minimize gender discriminatory language consist of: using gender neutral formulations when possible, and using a combination of masculine and feminine forms when not.

Amrhein et al. (2023) propose a novel approach to the gender-rewriting task. While the previous models rely on forward augmentation, this approach relies on *backward augmentation* and round-trip

translation to create a parallel dataset. This approach is used to create an English rewriter that matches or outperforms the results of Vanmassenhove et al. (2021). The backward augmentation approach consists of retrieving gender-fair data from large monolingual corpora and creating a rule-based pipeline to derive artificially biased text. The round-trip translation approach relies on the fact that most current machine translation models are socially biased. This can be exploited by using a biased model to translate from gender-fair text to a pivot language. This output is then translated back to the original language, creating a biased version of the original gender-fair text. The authors use the round-trip translation method to create a German rewriter, using English as a pivot language. Amrhein et al. (2023) resort to LLMs in order to generate additional gender-fair examples. We address this approach as future work in Section 7.2.

The only system that we were able to find for Romance languages was developed by Bellandi and Siccardi (2022) for the Italian language. While not being a "rewriter" (the system does not offer possible solutions to the gender discriminatory language it finds), we include it in this section due to its identification capabilities, since identifying non-gender-inclusive language is a core component of rewriter systems. They categorize two possible ways in which language can be gender discriminatory: when sentences contain "only the male form of a noun having a different female form", and "sentences containing nouns having the same male and female form, without any other grammatical element to stress reference to both genders". This last "problem" is also common in Portuguese (e.g. using *os docentes* with the male article *os*, when *docentes* is both the male and female term for "teacher"). A neural model was trained to recognize these situations, assigning a label to each.

Alhafni et al. (2022) developed a gender rewriter for Arabic, combining both rule-based and neural models. In this system, the rewriting is done only for the first and second grammatical persons (equivalent to *I* and *you* in English). The authors used the Arabic Parallel Gender Corpus (APGC), which contains "gender annotations and gender rewritten alternatives of sentences selected from OpenSubtitles 2018", as well as the English parallel. The APGC v2.0 (the version used during the study) also contains gender labels for each word specifying whether the word is gendered or not, female or male, and referring to the first-person or the second-person. The study presents three different models for rewriting with gender alternatives:

- A corpus-based model (CorpusR), that performs a lookup on the APGC.
- A morphological model (MorphR), using an analyser and generator for Arabic.
- A neural model: a character-level encoder-decoder model with attention, with the encoder being modeled as a two-layer bidirectional GRU and the decoder being modeled as a two-layer GRU with attention. The target gender label is appended to the input words as a special token, the expectation being that the model pays attention to the label in order to output the correct gender

alternatives.

A multi-step model that used all the rewriting components presented above was found to be the best model, achieving a BLEU of 98.92 on the test set.

This model is revisited in the work by the same authors (Alhafni et al., 2023), where they propose a web-interface where the users can specify the desired target genders (for first-person and second-person). The tool takes Arabic or English text as input, and outputs gender rewritten sentences according to the users' previously specified gender preferences.

3.1.3 Limitations

The models that we have presented all contribute in some way to our own work. However, we face challenges related to both the specificities of the Portuguese language and to the purposes of our work.

The English models (Vanmassenhove et al., 2021; Sun et al., 2021) are in some ways very similar to what we propose to implement for the Portuguese language, but there are some obvious limitations:

- Portuguese does not have a well-established gender-neutral alternative (such as the English *they*), as there are not one, but several different neopronouns in use today.
- In Portuguese, nouns and adjectives are gendered. Therefore, a rule-based neutral rewriter requires more than a POS-tagger and a dependency parser. We wish to translate "alunos" to "estudantes" or "alunes", but "students" is already a gender-neutral term and does not need to be translated.
- Since we propose a model where the user can select the pronouns for any named human entity, a co-reference resolution module is necessary. For example, if "João" uses *he/him* or *ele/dele* pronouns, we do not wish to translate "João estava entusiasmado" to "João estava entusiasmade".

The work of Amrhein et al. (2023) is the first to exploit the biases of machine translation models to generate gender-neutral examples, and to make use of Large Language Models (LLMs) to expand their gender-neutral data. However, we identify two issues with the round-trip translation approach when developing a rewriter for the Portuguese language:

- The lack of very large Portuguese monolingual datasets containing gender-fair language;
- The lack of consistency of the existing gender-fair data regarding the usage and choice of neopronouns: due to the diversity of Portuguese gender-neutral language proposals and neopronouns (detailed in Section 2.3), it is often the case that real examples of gender-neutral language are not consistent in terms of gender agreement. Examples of this phenomenon are depicted in Table A.1, in Appendix A.1.

Both the works of Diesner-Mayer and Seidel (2022) (for the German language) and Alhafni et al. (2023) (for the Arabic language) suffer from the same limitations:

- Even though they are built for gendered languages, they do not allow for gender-neutral rewriting using non-gender-specific expressions or neopronouns.
- The models do not allow for specification of the pronouns of named human entities (the Arabic model being limited to the specification of the gender of first and second grammar persons).

The Italian model (Bellandi and Siccardi, 2022) does not propose suggestions or actual rewriting of the discriminatory language it identifies, and therefore is only of interest to us in the sense of identifying non-inclusive language in Romance languages.

3.2 Gender Bias

The topic of gender bias in NLP systems is adjacent to our work, as it is inevitably related to gender inclusivity. However, an extensive analysis of the causes of gender biases in NLP and how to mitigate them is outside the scope of our work. Therefore, in this section, we present selected studies that might alert to possible ways in which gender biases can impact our own models. We are particularly interested in the lack of pronoun representation and variety in current models, and how gender imbalance in datasets may lead to biases and errors in machine translation and CR systems.

Before we delve into specific topics on gender bias, we would like to note that most of the studies that we present in this section have a strong focus on the English language. However, gender bias can manifest in different ways in Romance languages, and consequently in Portuguese. A form of gender bias in English might manifest in, for example, "the doctor" being tagged as a "male" even though there is no preceding pronoun to justify that gender choice. This phenomenon would not manifest in Portuguese (or any gendered language), because "a médica" is already "tagged" as "female" via the -a termination. Of course, this does not entail that word embeddings for gendered languages are not biased. Zhou et al. (2019) studied how gender bias manifests in gendered languages. The authors note that "when we align Spanish (ES) embeddings to English embeddings, the word "abogado" (male lawyer) is closer to "lawyer" than "abogada" (female lawyer)".

We would also like to note that concerns that are raised by most of the studies that we present in this section are largely focused on binary gender representation. According to Dev et al. (2021) one of the reasons for such is a *dataset skew*: the large text dumps the language models tend to use as data have "severe skews with respect to gender and gender-related concepts". For example, English Wikipedia has over 15 million mentions of *he*, 4.8 million of *she* and 4.9 million of *they*. The mentions of

they are usually plural *they* and not the gender-neutral singular pronoun, and mentions of neopronouns are usually not meaningful.

3.2.1 Modelling Pronouns for Gender Fairness

Current NLP systems have a notorious difficulty in processing pronouns in a gender-fair way. In systems focused on the English language, the bias against *her* and the singular *they* is amplified, even in current state-of-the-art systems (Munro and Morrison, 2020).

The 2022 study of Brandl et al. gives us some insight into how existing models process neopronouns. In one of their experiments, the authors studied the performance of state-of-the-art NLP models on Natural Language Inference (NLI) and Coreference Resolution (CR) tasks. The dataset used included the pronouns *he*, *she*, the singular *they*, and the neopronoun *it*. For NLI, there was a "very small drop in performance for the datasets with gender neutral pronouns compared to the original sentences". Two models were used for the task: mBERT, the multilingual version of BERT (Devlin et al., 2019) (a language representation model that consists of a multi-layer bidirectional Transformer (Vaswani et al., 2017) encoder); and XLM-R (Conneau et al., 2020), a cross-lingual model that uses self-supervised training techniques. For mBERT, the performance drop when using gender-neutral pronouns was 0.09 – 1.51%. For XLM-R the drop was 0.21 – 4.71%. The CR experiment ran on the NeuralCoref 4.0 in spaCy (neu), using the English Winogender (Rudinger et al., 2018) dataset. The results showed a "drop in performance from gendered pronouns (she, he) to both gender-neutral pronouns (they, xe)". The accuracy for *xe* was 0%.

Different methods of modelling pronouns might contribute to mitigating the lack of gender/pronoun representation. Lauscher et al. (2022) define five desiderata for modelling pronouns in NLP systems:

- A model should not assume an individual's pronouns.
- A model should be capable of processing not only the "standard set of pronouns in a language", but also neopronouns.
- A model should allow for the addition of new pronouns.
- A model should "allow for multiple, alternating, and changing pronouns", due to the pronouns that an individual uses being subject to change over time.
- A model should "provide an option to set up individuals' sets of pronouns".

The authors design an experiment where they follow a *delexicalization* strategy for modelling pronouns, where "the model learns a single representation for all pronouns and relies on other task-related conceptual and commonsense information for disambiguation". With this method, the goal would be for

the model to not learn lexical cues from context (for example, a specific pronoun being associated with a certain proper name). Two variants of the original dataset were created: one where all pronouns in the test set were replaced with the respective POS token, and one where pronouns were replaced on the train, dev, and test splits. The chosen model was RoBERTa (Liu et al., 2019) (a BERT variant with an improved pretraining procedure that includes training the model longer and over more data) as the base encoder, and the selected task was CR. The results for the variant where pronouns were only replaced in the test set dropped on average by 21.2% in F1-measure. According to the authors, this demonstrates a "heavy reliance" of the model on lexical cues. The variant were pronouns where replaced in all splits resulted in a drop of -4.2 F1. The authors remark that all pronouns were replaced, including non-third person pronouns, and that they could expect "even smaller drops from a more careful selection of replacements". As such, it is possible for systems to maintain high performance while modelling pronouns using different paradigms.

3.2.2 Bias in Machine Translation

Analysing gender bias in machine translation systems is particularly important, due to the possibility of bias causing incorrect translations. The differences in grammatical gender systems across languages are an issue when it comes to neural machine translation. We have previously seen an example of English-Arabic translation in Section 3.1, with the work of Alhafni et al. (2023).

A possible way to mitigate bias in machine translation from non-gendered to gendered languages is the introduction of specific gender tags into models. Saunders et al. (2020) propose a method of incorporating gender tags into the model for "translating coreference sentences where the reference gender label is known". Of particular interest to our work is their implementation of a gender-neutral tag, although non-specific to the target language. To provide an example given by the authors, the English sentence "the trainer finished their work" is translated to Spanish as "DEF entrenadorW_ END terminó su trabajo", where DEF and W_EN are a non-gender specific placeholder article and a noun inflection, respectively. For English to Spanish translation, the accuracy of the best model, trained on the WinoMT corpus — a publicly available challenge set and evaluation protocol for the analysis of gender bias in MT designed by Stanovsky et al. (2019) — was 56.5%, compared to a baseline of 4.2%.

There are several ways to implement these gender tags. A study from Vanmassenhove et al. (2018) explored a method of incorporating gender information in machine translation systems. The authors tagged a parallel corpus of language pairs, containing both gendered and non-gendered languages, with information on the speaker's gender. The machine translation models tagged with gender information improved over the baseline model in French, Italian, and Danish, with statistical significance ($p < 0.05$). Interestingly, given the differences in the manner of speech between men and women, the models tagged with gender tend to prefer the terms most used by the gender in question. As such, as the

authors remark, “even for languages that do not mark gender overtly (i.e. grammatically), it can still be beneficial to take the gender of the author/speaker into account”.

Cho et al. (2019) proposed a schema for evaluating gender bias in a Korean-English machine translation system. Korean and English are interesting languages to study translation due to both having gender-neutral pronouns. The test set consisted of sentences in which gender neutrality should be maintained. Sentences categorized as “formal” were more biased towards male than “informal” sentences. Sentences categorized as “occupation” reflected the most bias: in Google Translator[1] mentions of “engineers”, “technicians”, and “professors” were significantly assumed to be male in translation to English. The South Korean machine translation cloud service Naver Papago[2], however, showed the opposite bias: “researchers” and “engineers” were assumed to be female in translation. The authors note that this is probably due to a team effort to reduce social biases in the system, and remark that the final objective should be the preservation of the gender neutrality, and not a “half-half guess” of a gender binary.

3.2.3 Bias in Coreference Resolution

The work of Cao and Daumé (2021) is an obligatory mention for understanding not only how gender biases can impact CR models, but also how the concept of gender is treated across all stages of NLP systems. The authors categorize several ways in which bias can enter the “machine learning lifecycle of coreference resolution systems”, of which: in the task definition for annotations (of the thirteen English datasets annotated for CR analysed by the authors, none of the annotation guidelines included neopronouns, and *he* occurred more than twice as frequently as all other pronouns); in data input, where bias can arise from the selection of texts to use as data; and model definition, where bias can arise from external resources, chosen features, etc. The authors also remark on the assumptions about social gender that are made by the NLP community. Notably, of 22 papers on coreference, 5.5% distinguish linguistic from social gender, 94.4% assume that gender is binary, and only one paper allows for gender neutral pronouns (singular *they* or neopronouns). The authors also introduce a new dataset (**GICoref** [3]) for “evaluating current coreference resolution systems in the contexts where a broader range of gender identities are reflected”, consisting on 95 documents.

State-of-the-art CR models for the English language have notorious difficulty in recognizing gender-neutral pronouns and processing them accordingly. However, there have been recent efforts to evaluate current CR systems for gender inclusivity. Zhao et al. (2018) developed a challenge corpus for evaluating gender bias in CR models, WinoBias. Rudinger et al. (2018) created a “Winograd schema-style set of minimal pair sentences that differ only by pronoun gender” to evaluate gender bias in CR systems.

[1] https://queogle.com Accessed on 01-12-2022
[2] https://papago.naver.com/ Accessed on 01-12-2022
[3] https://github.com/TristaCao/into_inclusivecoref Accessed on 05-12-2022

The authors correlate gender predictions from three CR models with real-world statistics on gender and occupations. An example of discrepancy noted in the study is the occupation "manager", which is 38.5% female in the U.S. (U.S. Bureau of Labor Statistics), but no managers are predicted to be female by the analysed systems. The authors also concluded that the three analysed systems (rule-based, statistical, and neural models) did not "behave in a gender-neutral fashion", meaning that they "exhibit sensitivity to pronoun gender" even though the test sentences where pronoun resolution was not gender dependent, as validated by human annotators.

3.3 Tools & Models

3.3.1 Portuguese Wordnets

A Wordnet is a lexical database of nouns, verbs, adjectives, and adverbs. Words are grouped into *synsets* (cognitive synonyms), each expressing a distinct concept. Synsets are interlinked by semantic and lexical relations, such as synonym-antonym and hyponym-hyperonym. The first Wordnet was developed for the English language at Princeton (Miller, 1995), but variations for different languages have since been developed[4]. Several wordnets have been developed for the Portuguese language. Here we present a few of the most noteworthy.

OpenWordnet-PT (de Paiva et al., 2012) (or OpenWordnet-PT (OWN-PT)) is an open-access wordnet for Portuguese that follows the mappings of the Princeton Wordnet (University, 2010).

ONTO-PT (Gonçalo Oliveira and Gomes, 2014) is the largest Portuguese Wordnet as of date. It was built with a completely automatic approach, and contains semantic relations that are not present in the original Princeton Wordnet.

Wordnet.PT (Marrafa, 2002) is the earliest Portuguese Wordnet we have knowledge of, being in development since 1999. It was been extended to accommodate other Portuguese variants since 2011 Marrafa et al. (2011). However, it is currently a closed wordnet.

3.3.2 NLP Resources for Portuguese

Several tools for preprocessing Portuguese texts and performing NLP tasks are currently available. Here we present only the most commonly used and/or the ones that we have used during the development of our models and that can be used in future work.

[4] http://globalwordnet.org/resources/wordnets-in-the-world/ Accessed on 20-11-2022

NLTK (Bird et al., 2009) is a suite of Python modules that includes models for processing tasks such as segmentation and tokenization, stemming, POS tagging, stemming, stopword removal, and simple concordancing. NLTK offers corpora from the Floresta Sintá(c)tica treebank[5].

spaCy is a Python library that makes available trained pipelines for Portuguese. It uses the Universal Dependencies (UD) Portuguese treebank Bosque (Rademaker et al., 2017), which is part of the Floresta Sintá(c)tica.

Stanza (Qi et al., 2020) is a Python package for natural language analysis. Like spaCy, it also makes use of Bosque. It provides pre-trained NLP models trained on the Bosque treebank, that cover tokenization, multi-word token (MWT) expansion, lemmatization, POS and morphological features tagging, and dependency parsing. Stanza also provides a Portuguese model for constituency parsing.

STRING is a hybrid statistical and rule-based natural language processing chain for Portuguese (Mamede et al., 2012). STRING performs preprocessing (text segmentation, tokenization, and POS tagging), lexical analysis, statistical and rule-based POS disambiguation, and dependency parsing.

3.3.3 Large Multilingual Machine Translation Models

As we have previously established, one of the objectives of this work is to create a neural model that rewrites text to be as gender-neutral or gender inclusive as possible. This task has many similarities to machine translation, due to both being sequential tasks. While an in-depth study of neural machine translation technologies is out of the scope of our work, here we provide a brief overview of the state-of-the-art. Of particular interest to us are the topics of Neural Machine Translation (NMT) for languages with low NLP resources, due to the likelihood of our work developing a relatively small corpus, and translation from non-gendered to gendered languages. The way in which existing models solve this last issue often provides some insight into "gender-neutralization" of gendered languages inside the context of NLP and not social linguistics, and is therefore interesting to us.

Defining what constitutes a low-resource language is somewhat complex, and usually entails quantifying resources (labeled and unlabeled), pre-existing tools for processing that particular language, and the effort that the NLP research community is making to investigate processing for that particular language. Joshi et al. (2020) divide languages into six classes of different levels of resource availability. Where Portuguese features into these classes is up for debate. While it is sometimes considered a low-resource language (Przystupa and Abdul-Mageed, 2019), and over the course of this work we remark the lack of resources for this language when compared to a very high-resource language (e.g. English),

[5]https://www.linguateca.pt/ Accessed on 20-11-2022

we would like to note that 88.17% of languages (amounting to one billion speakers) belong to class 0 of the proposed language classes of Joshi et al, meaning that they are "ignored in the aspect of language technologies" and have extremely limited resources. This is not the case for Portuguese, as can be understood by the number of NLP tools we have presented in this very section.

We follow with a brief overview of some widely used multilingual machine translation models.

M2M-100 (Fan et al., 2021) is a non-English centric Transformer-based model which can translate from and between 100 languages, without pivoting to English. Although the encoder and decoder are shared between languages, the model possesses a language-specific layer. The languages are grouped based on their vocabulary and the amount of training data.

NLLB-200 (Costa-jussà et al., 2022) is a Transformer-based model which can translate 200 different languages. It was trained on data obtained with data mining techniques tailored for low-resource languages. It uses a language specific encoder (LASER[6]) , which has also been made open-source.

MBART (Liu et al., 2020) is a sequence-to-sequence denoising auto-encoder, which is primarily intended for the task of machine translation.

[6]https://github.com/facebookresearch/LASER Accessed on 05-09-2023

4

Datasets

Contents

In this chapter, we present the corpora we have used while compiling datasets for training our neural gender-neutral rewriter model and evaluating our systems.

We chose to split our dataset into five text categories to analyze the performance of our models in different types of text, and how they account (or not) for variability of sentence structure and vocabulary. To provide us with a better understanding of the differences between these text categories, we have calculated metrics regarding the complexity of each type of text. This analysis is detailed in Section 4.3. The five text categories we have designed are described in further detail below:

Literary Texts Selected works found in the *DIP* collection[1] from Linguateca[2] resource center. DIP is a shared task whose goal is to identify characters and respective attributes in literary works (Santos et al., 2022). In order to avoid examples with an orthography that might be too different from modern Portuguese, we have only selected works released after 1910. We used NLTK to tokenize the raw data at a sentence level. Cleaning and processing the data consisted of the removal of tags (such as chapter indications, language tags, etc.) and author notes. Literary texts tend to consist of sentences with varying structures and rich vocabulary.

Journalistic texts Random sample of sentences found in the *NaturaPublico94* dataset, from Projecto Natura[3]. The original corpus contains the first 2 paragraphs of each article in the Portuguese newspaper *Público*, retrieved during the period of 1991 to 1994. *NaturaPublico94* was retrieved during 1994, being the most recent newspaper. The raw texts were tokenized at a sentence level using NLTK. Sentences that were mis-tokenized or deemed too short, meaning that they consisted only of one or two characters long or only one word, were removed. Adding journalistic texts to our dataset allows us to enrich the data with formal vocabulary and several named entities.

Dialogues Random sample of extracted sentences from the *SubTle* corpus (Ameixa and Coheur, 2013; Ameixa et al., 2014). *SubTle* aggregates dialogues from movie subtitles, extracted from IMDB[4] and pertaining to one of four movie genres: Horror, Scifi, Western, and Romance. During the cleaning of the dataset, we removed speaker tags. Although these dialogues can be considered synthetic, since they do not originate from real-life conversations, they provide us an insight on the performance of our models in direct speech.

Social Media Random sample of tweets from the *Portuguese Tweets for Sentiment Analysis*[5] dataset, which contains examples retrieved mainly from 01/08/2018 to 20/10/2018. We only used tweets from

[1] https://www.linguateca.pt/aval_conjunta/dip/colecao.html
[2] https://www.linguateca.pt/
[3] https://natura.di.uminho.pt/ jj/pln/corpora/
[4] https://www.imdb.com/
[5] https://www.kaggle.com/datasets/augustop/portuguese-tweets-for-sentiment-analysis

the “no theme” partition of the dataset. We removed links, mentions, hashtags, and emojis. Social media comments and posts tend to include noise (in the form of misspelled words, emojis, and strange characters). However, it might be worth to note that the majority of written examples of Portuguese gender-neutral language are found in social media posts and comments. As such, we believe training and evaluating our models’ performance in social media examples to be essential.

Simple Sentences Samples from the Portuguese dataset of the Tatoeba (Tiedemann, 2020) corpus. Tatoeba is a multilingual data set of machine translation benchmarks derived from user-contributed translations. It consists of relatively simpler (both in terms of vocabulary and syntactic construction) and less noisy sentences.

4.1 Automatically Curated Set

We curated sets of 5,000 sentences from each text category, amounting to a total of 25,000 examples.

The gender-neutral alternatives of each example are generated by the rule-based model. This parallel dataset, containing the original (binary-gendered) sentences and the respective gender-neutral version, was used for training our neural model. Table 4.1 depicts one example for each dataset category.

The automatically curated set is composed of approximately 60% gendered (14874 sentences) and 40% non-gendered sentences (10126 sentences). For our purposes, we consider a sentence to be *gendered* if it contains proper nouns, personal pronouns, or human referents.

Category	Original Sentence	Gender-neutral Sentence
Literary	“É orgulhoso e de opinião, como ele só!”[6]	“É orgulhose e de opinião, como elu só!”
Journalistic	“Fomos à procura deles e organizámos um almoço comemorativo.”[7]	“Fomos à procura delus e organizámos um almoço comemorativo.”
Dialogue	“Precisa saber só de olhar para a mulher, sem ela dizer.”[8]	“Precisa saber só de olhar para a pessoa, sem elu dizer.”
Social Media	“foi ela quem fez o exorcismo.”[9]	“foi elu quem fez o exorcismo.”
Simple Sentences	“Eu estou viciado em mascar chiclete.”[10]	“Eu estou viciade em mascar chiclete.”

Table 4.1: Dataset categories and respective examples. The original sentences contain idiomatic expressions, which we tried to capture in the English translation.

4.2 Manually Curated Test Set

For curating our test set, we selected an additional 100 sentences from each of the five text categories we have previously described. All 500 examples are *gendered*, containing either named entities, human referents, or personal pronouns.

We have manually annotated the 500 sentences in the collection. A sample of 100 examples belonging to the collection was annotated by other 5 fellow researchers in order to calculate the annotator agreement. The annotators followed an annotation guide for the *elu* system, whose rules are consistent with the ones employed in our rule-based model and are inspired by the proposals presented in Section 2.3, particularly the ones authored by Caê (2020) and Santos and Marques (2021). The full guide can be found in Appendix B.

We calculate the agreement using the metrics Word Error Rate (WER), Character Error Rate (CER) (Morris et al., 2004), and Exact Match. We achieve a WER of **2.15%**, a CER of **0.54%**, and an Exact Match score of **82%**. We assume that these results reflect the quality of the full collection. The reasoning behind the usage of these metrics is the following:

- We use WER and CER as metrics for evaluating our gender-neutral rewriting models, as detailed in Section 6.1.

- This test set is composed of curated sentences, and as such do not contain extra/missing spaces or strange characters. Therefore, we can use Exact Match as a metric for evaluating if two annotators are in full agreement regarding the rewriting of a sentence (not just the rewriting of a single word).

Annotation disagreements often arise either from the existence of several possible gender-neutral alternatives, or from uncertainty over if a certain term should be neutralized. Examples of these types of disagreements are found in Table 4.2. In the first sentence, the devil (*diabo*) may be considered a genderless entity, and therefore terms related to it should not be rewritten. However, the female form of *diabo*, *diaba*, can be used, which may be used as an argument in favor of *diabo* being a gendered term. In sentence 2, the genderless term *doentes* is a synonym to the gender-neutral term *enfermes*, derived from the term *enfermos/enfermas*.

[6] English: “He is proud and opinionated, as only he can be!”
[7] English: “We went looking for them and organized a celebratory lunch.”
[8] English: “He needs to know just by looking at the woman, without her saying so.”
[9] English: “she was the one who performed the exorcism.”
[10] English: “I’m addicted to chewing gum.”
[11] English: “The devil is on the loose.”
[12] English: “the sick were often exposed in the street.”

Annotator X	Annotator Y
Está **o diabo** à solta.[11]	Está **ê diabe** à solta.
Ês **enfermes** eram, muitas vezes, expostes na rua.[12]	Ês **doentes** eram, muitas vezes, expostes na rua.

Table 4.2: Annotator disagreements are marked in bold.

4.3 Data Complexity and Structure

We have previously stated that our datasets differ in terms of complexity and structure. Langacker (1973) defines a *complex sentence* as one that consists of more than one clause, and a *clause* as a sentence constituent that contains a verbal element of some kind and can (with slight modifications), stand alone as a sentence.

Having these definitions in mind, we have calculated:

- *The average sentence length of our data examples*, depicted in Figure 4.1.

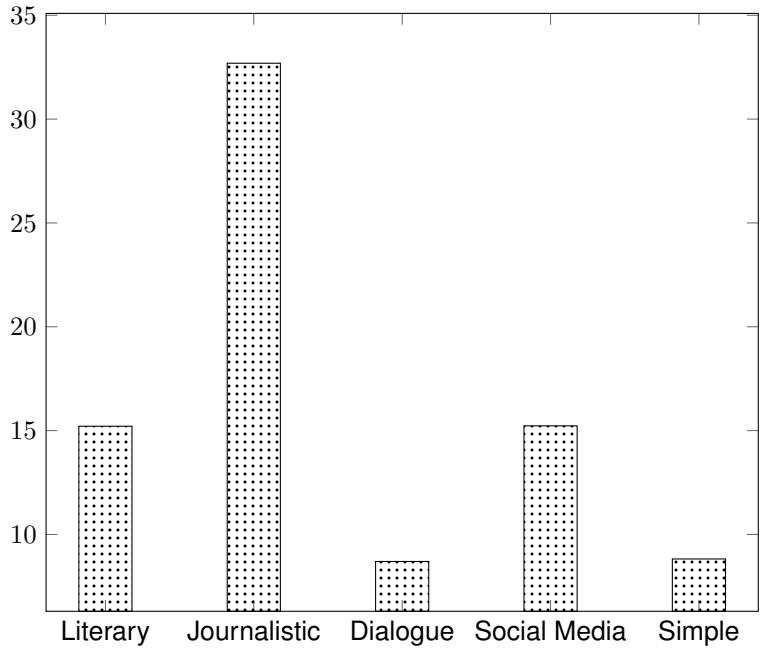

Figure 4.1: Average sentence length (word-wise).

- *The average number of verbal phrases for each category*, depicted in Figure 4.2. For analysing clauses, we have used the Portuguese constituency parser made available by Stanza Qi et al. (2020). This parser has been trained on the CINTIL-TreeBank dataset (Branco et al., 2011), a corpus of syntactic constituency trees of Portuguese texts. In CINTIL's syntactic constituents representation, an S can be projected out of a VP in case this VP's head has an internal complement.

This can lead to an S constituent having as children another S node and a PUNCT node, as depicted in Figures 4.3 and 4.4. As such, we have determined that assuming each S constituent represents one clause is not correct, and counting instead each *verbal phrase* as an indicator of a new clause to be more accurate. Figure 4.3 depicts the constituency parsing tree of a sentence belonging to the Simple category, which is the category with the lowest average number of verbal phrases. Figure 4.4 depicts the tree of a sentence belonging to the Journalistic category, which is the category with the highest average number of verbal phrases.

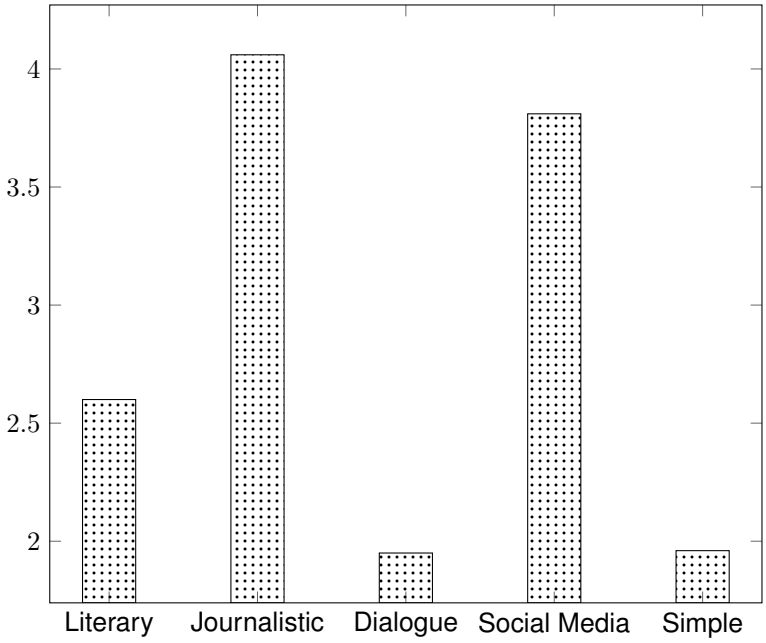

Figure 4.2: Average number of verbal phrases in a sentence.

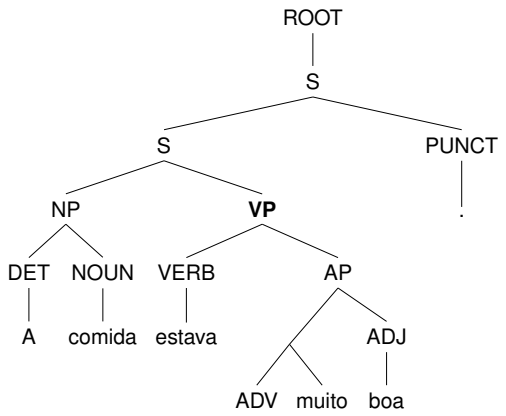

Figure 4.3: Average sentence length (word-wise).

- *The number of words in the vocabulary of each dataset*, depicted in Figure 4.5.

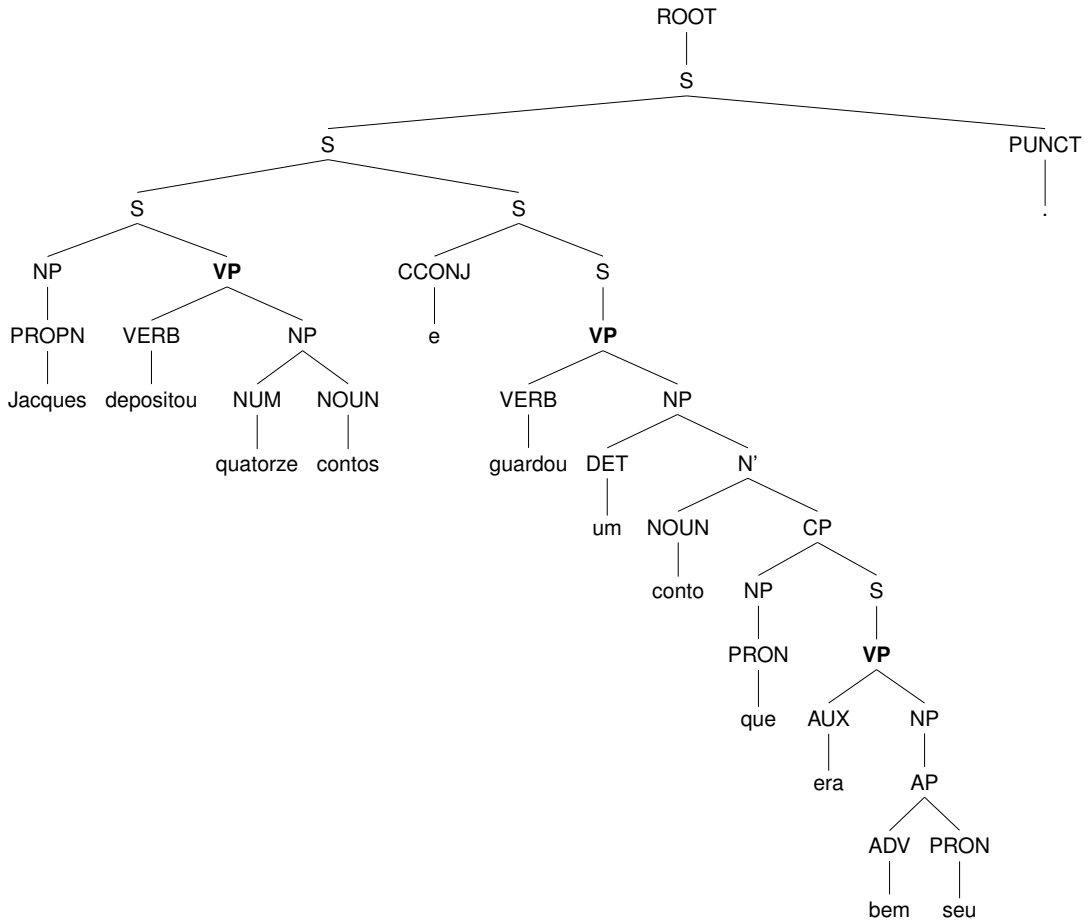

Figure 4.4: Average sentence length (word-wise).

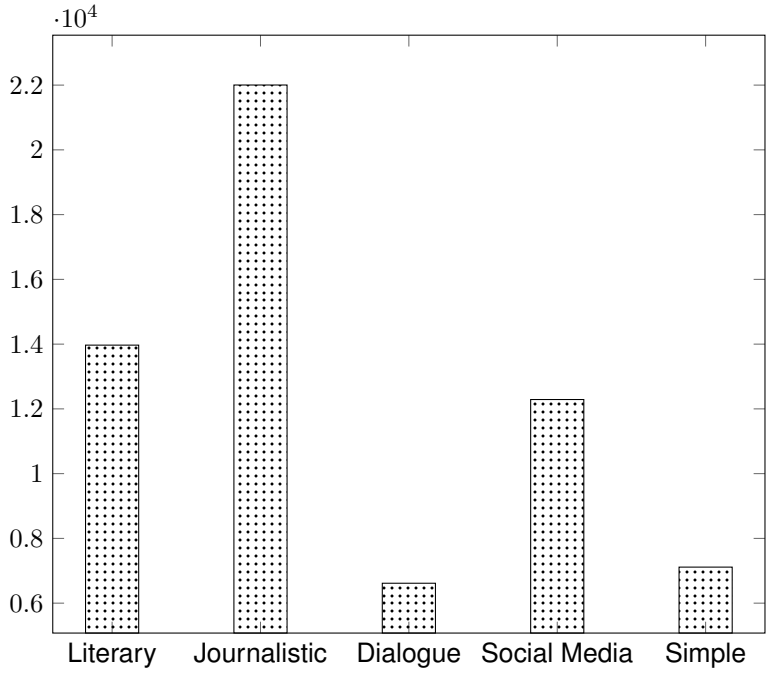

Figure 4.5: Vocabulary of each dataset category.

As can be inferred by these metrics, the Journalistic text category is the one with the most complex sentences, and the Simple category is the one with the least complex sentences. In most categories (except Social Media), if sentences tend to be remarkably long, then there also tends to be a relatively high number of verbal phrases and a long vocabulary. However, the Social Media category sentences are of average length (around 15 words) and vocabulary, but hold almost as many verbal phrases as the Journalistic category, the most complex. The sentences belonging to the Dialogue category are, according to these metrics, of similarly low complexity as the sentences belonging to the Simple category.

5

Models

Contents

As we have previously discussed, our work aligns itself with the category of *gender-neutral rewriter systems*. In this chapter, we introduce the architecture of our rule-based model for gender-neutral rewriting, analyzing and discussing the pipeline in detail. Furthermore, we describe the setup and fine-tuning method of our neural model.

5.1 Rule-based Model

5.1.1 Overview

As depicted in Figure 5.1, the Rule-Based Model (RBM) is composed of three main modules.

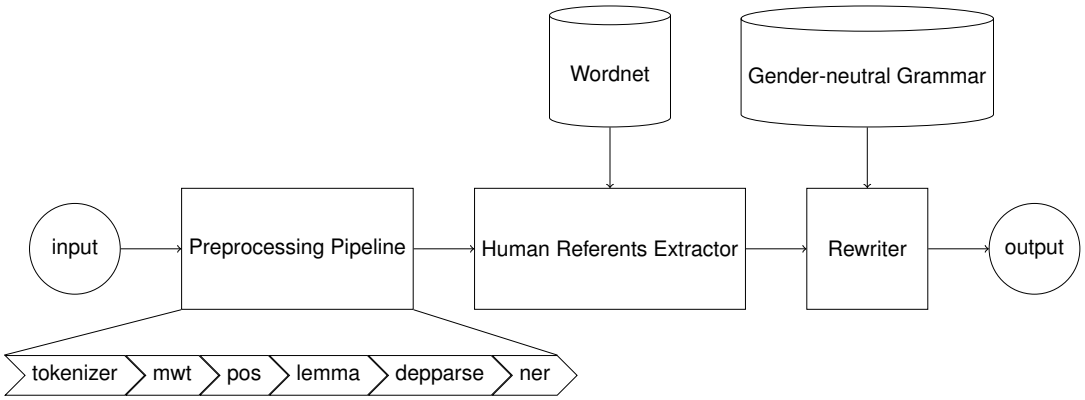

Figure 5.1: RBM Pipeline.

- A Stanza neural pipeline for preprocessing, tagging, and parsing of input, described in detail in Section 5.1.2.
- An extractor module that stores information on proper nouns, nouns that refer to people, and their respective heads in the sentence. We call this module Human Referents Extractor, and refer to it throughout the text as the extractor module. It is described in detail in Section 5.1.3.
- A rewriter module that, given information provided by the dependency parsing graph, the grammar and database detailed in the Appendix A, and information from the extractor module, identifies binary-gendered terms that refer to people and attempts to rewrite them as gender-neutral. It is described in detail in Section 5.1.4.

The RBM functions at a sentence level, meaning that word dependencies are only analyzed in the context of a single sentence. At the time of writing, this is the case for all gender-neutral rewriters (we have presented these in more detail in Section 3.1). While this has no impact on the task of gender-neutral rewriting, since every gendered term becomes gender-neutral without regard for the human

entities' preferred pronouns, it may have an impact if we desire, for instance, to replace the pronouns of a specific human entity. In those cases, a coreference resolution module that correctly processes gender-neutral language and neopronouns would be necessary. We revisit this topic in future work.

The RBM considers as binary-gendered (and therefore in need to be rewritten):

- Any usage of nouns that refer to people in its binary-gendered way (e.g. *aluno*, *aluna*, or their plural forms *alunos* and *alunas*). "Nouns that refer to people" includes proper nouns.
- Terms that are subject to gender agreement with nouns that refer to people (adjectives, determiners, verbs, and pronouns).

The RBM allows for users to decide whether they wish for determiners that precede proper nouns to be omitted (i.e. "O Fred" becomes "Fred"), as depicted in Figure 5.3. The model also allows users to select whether they wish to check the list for already existing gender-neutral expressions, as depicted in Figure 5.2.

The RBM has been developed in Python and, at the time of writing, its user interface consists of a simple script. It has been made open-source[1]. The instructions given to the user when first running the script are depicted below.

```
1  Welcome to Gender Neutralizer!
2
3  Gender Neutralizer assumes that the input text is written in a
4  binary-gendered portuguese.
5  It will attempt to replace the pronouns of any binary-gendered entity with a
6  desired  gender neutral form.
7
8  Currently, Gender Neutralizer only supports a neutral form with an -e termination.
9  Gender neutralizer uses the gender neutral neopronoun elu.
10
11 Do you wish to omit determinants that precede proper nouns?
12 This is recommended for legibility. (y/n)
13 y
14
15 Do you wish to check already existent gender-neutral alternatives to words? (y/n)
16 y
```

Listing 5.1: RBM user script.

[1] https://github.com/leonorv/pt-gender-neutralizer/

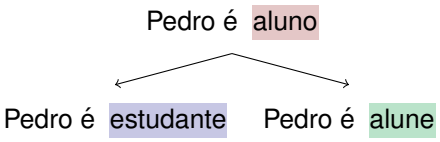

Figure 5.2: RBM outputs: the sentence on left is generated if the user wishes to check already existing gender-neutral expressions (generating "estudante"); the sentence on the right is generated otherwise (generating "alune").

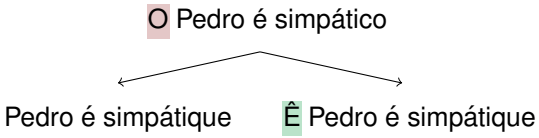

Figure 5.3: RBM outputs: the sentence on left is generated if the user wishes to omit determiners that preced a proper noun; the sentence on the right is generated otherwise.

We follow with a set of example sentences processed by RBM. The processing included both the omission of determiners that precede proper nouns, and rewriting with already existing gender-neutral expressions.

(1) **Input:** O Fred é bom cozinheiro e decidiu fazer uma tarte para o seu marido.[2]

Output: Fred é boe cozinheire e decidiu fazer uma tarte para sue cônjuge.

(2) **Input:** Ele saiu de casa para ir ao mercado, mas assim que chegou lá viu que os vendedores estavam em greve.[3]

Output: Éle saiu de casa para ir ao mercado, mas assim que chegou lá viu que ês vendedores estavam em greve.

(3) **Input:** O seu marido ficou desiludido e acabaram por ir jantar fora.[4]

Output: Sue cônjuge ficou desiludide e acabaram por ir jantar fora.

5.1.2 Preprocessing Pipeline

The preprocessing pipeline consists of tokenization, POS-tagging, dependency parsing, and named entity recognition models made available by the Stanza (Qi et al., 2020) toolkit.

At the time of writing, Stanza does not make available any Portuguese named entity recognition model. Therefore, the preprocessing pipeline currently makes use of the Stanza named entity recognition model for Spanish, chosen due to the similarity between the two languages.

[2] English: "Fred is a good cook and decided to make a pie for his husband."
[3] English: "He left home to go to the market, but as soon as he got there he saw that the salesmen were on strike."
[4] English: "His husband was disappointed and they ended up going out to dinner."

The tools for preprocessing Portuguese text that we have presented in Section 3.3.2 differ in terms of the features that they capture, and their performances differ according to which NLP task they are being used to solve (Gonçalves et al., 2021). For our purposes, we chose the tool that had POS tags that best suited our needs. Stanza (Qi et al., 2020) was chosen as our preprocessing tool, as it uses the well-documented Universal Dependencies POSs tags[5] and morphological features[6], such as plurality and gender.

Our Stanza neural pipeline contains five processors:

- **tokenize:** Segments the input document into sentences, each containing a list of tokens.
- **mwt (Multi-Word Tokens):** Expands tokens into multiple syntactic words (necessary to create the Universal Dependencies correctly).
- **pos:** Labels tokens with their Universal Dependencies POSs tags (UPOS[7]), treebank-specific POSs tags (XPOS), and universal morphological features UFeats[8].
- **lemma:** Performs lemmatization on words.
- **depparse:** Determines the syntactic head of each word in a sentence and its dependency relation. The result of applying it to a full sentence is exemplified in Figure 5.4.

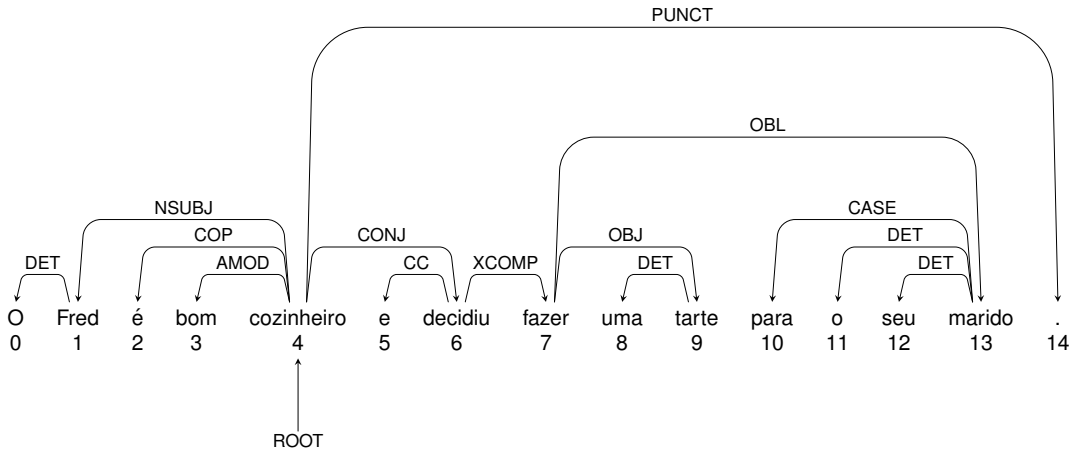

Figure 5.4: Dependency Parsing for Sentence 1 as provided by Stanza

5.1.3 Human Referents Extractor

The extractor module sweeps the input text to find all proper nouns, nouns, and pronouns that refer to people. Their positions in the text, as well as the positions of their heads in the dependency parsing

[5]https://universaldependencies.org/u/pos/ Accessed on 05-12-2022
[6]https://universaldependencies.org/u/feat/index.html Accessed on 05-12-2022
[7]https://universaldependencies.org/u/pos/
[8]https://universaldependencies.org/u/feat/index.html

graph are then stored and used in the rewriter module.

Since this task requires access to a Portuguese wordnet (Miller, 1995), we settled on using OpenWordnet-PT (de Paiva et al., 2012). This decision is grounded on three of its characteristics:

- OWN-PT is open-source;
- It has been integrated with the Python library WN, as part of the Open Multilingual Wordnet (OMW) Collection (Goodman and Bond, 2021), and integrated with the Python library NLTK (Bird et al., 2009; Loper and Bird, 2004). The NLTK integration and documentation makes OWN-PT simpler to install and work with;
- It follows the mapping of Princeton Wordnet, whose semantic relations are well-documented.

The extractor module stores the indexes of terms that belong to one of the following categories:

Human Referents Words that are POS-tagged as proper nouns, as well as personal pronouns, are automatically considered to be referent to people. NLTK allows access to the lexicographer file of each word sense. Lexicographer files are split into 45 categories. Category 18 contains "nouns denoting people". Our extractor module checks if the lexicographer file name of the current synset corresponds to this category. If so, that noun is referent to a human, and, as such, terms related to it should have the correct gender form. After processing Sentence 1, the extractor module would mark `[1, 4, 13]` as tokens referring to people (corresponding to *Fred*, *cozinheiro*, and *marido*),

Heads of Human Referents Proceeding with the example of Sentence 1, `[4, 4, 7]` are marked as the heads of the tokens *Fred*, *cozinheiro*, and *marido*. This can be seen in Figure 5.4 where, for example, *fazer*, with index `7`, is a syntactic head of *marido*.

Proper Nouns Although proper nouns can be considered human referents, they have a particularities when it comes to gender agreement regarding determiners. The determiner that preceeds a proper noun can be omitted to maintain gender ambiguity. For example, "**O Fred** é inteligente."[9] becomes "**Fred** é inteligente"

Terms that have a gender-neutral alternative The extractor module checks a table with gender-neutral alternative terms and marks the terms that do have a respective alternative. The full table is depicted in Appendix A. It is worth to note that, although these alternatives may be gender-neutral, rewriting related terms to keep gender agreement may still be needed. For example, "**O homem** é inteligente."[10] becomes "**A pessoa** é inteligente."[11].

[9] English: "Fred is smart."
[10] English: "The man is smart."
[11] English: "The person is smart."

Terms which are already gender-neutral Terms which are already gender-neutral are marked, since rewriting them will not be needed. For example, "**A pessoa** é inteligente." should not be rewritten.

5.1.4 Rewriter

The final module of the RBM rewrites gendered terms related to human referents, whose positions in the sentence are stored by the extractor module.

In Portuguese, nouns, pronouns, clitic pronouns, determiners, adjectives, and verbs can be gendered. The rules used for "gender-neutralizing" each term depend on their word class, as described in further detail below.

Nouns & Pronouns Since the neopronoun "elu" is the most used in the third neutral gender proposal we have presented in Section 2.3, the system we employ in our model currently only uses that specific neopronoun. Certain gendered nouns have an existent gender-neutral synonym, which can be used instead of rewriting the gendered noun using rules. The rewriter module performs a lookup to a table (depicted in Appendix A.2) containing some of these gender-neutral terms and uses them accordingly.

Clitic Pronouns If a clitic pronoun is gendered, then it is rewritten (e.g. "Eu vou vê-**la**." → "Eu vou vê-**le**.").

Determiners Definite articles that precede a proper noun are omitted (e.g. "**O** João é feliz." → "João é feliz."). Other types of determiners are neutralized if their head in the dependency graph is referent to a human (e.g "João é **um** rapaz." → "João é **ume** jovem.").

Adjectives The task of checking whether an adjective refers to a certain referent is complex, since an adjective can be in multiple positions in a sentence. The rule-based rewriter module assumes that an adjective should be rewritten if either the adjective itself or its head in the dependency graph has been marked by the extractor module as either referent to a human or a head of a term marked as referent to a human. This rule tends to correctly rewrite adjectives in sentences with a relatively simple construction. However, it fails in sentences where an adjective and a human-referent noun share a root term in the dependency graph. This is illustrated in Figure 5.5 and Figure 5.6.

Verbs Most verb tenses are not gendered, with the exception of some main verbs that require an auxiliary verb, such as past participle forms (e.g "João foi levad**o** a jantar." → "João foi levad**e** a jantar."). The RBM rewrites verb forms that require an auxiliary verb different from the verb "ter" (to have), since that particular verb is used in several non-gendered tenses. We make an exception for gerund, which

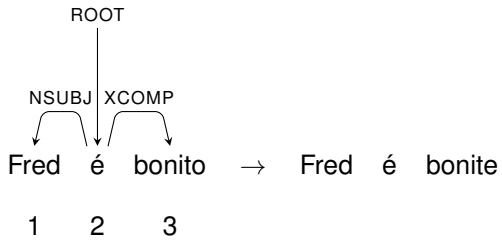

Figure 5.5: Since both the terms *Fred* and *bonito* share a head node, *é*, and *Fred* is a human referent, the adjective is rewritten.

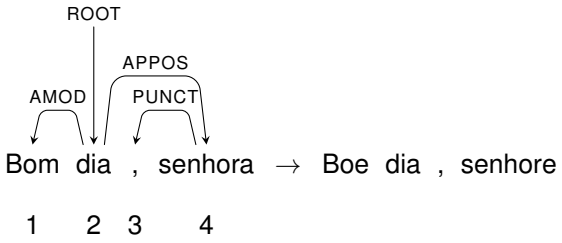

Figure 5.6: The same rule that worked for the case on Figure 5.5 fails with this sentence structure. Since *senhora* is a human referent, the adjective *bom* is incorrectly neutralized, even though it refers to the term *dia*.

can require an auxiliary verb different from "ter", but is not a gendered form (e.g "João foi andando." → "João foi andando.").

5.2 Neural Models

5.2.1 Base Models

We argue that we can look at the task of gender-neutral rewriting not only as a sequence-to-sequence task, but as an *intralingual translation* problem. This type of translation, also referred to as *rewording*, can be considered extremely peripheral to translation studies (Zethsen, 2009). As such, while designing our neural approaches, we have regarded gender-neutral rewriting as a machine translation problem.

As of writing, LLMs have been gaining traction, due to their ability to handle a variety of tasks. However, a 2023 analysis of the performance of LLMs (Zhu et al., 2023) shows that one of the currently used many-to-many multilingual translation models, NLLB-1.3B (Costa-jussà et al., 2022), still outperforms the best instruction-tuned LLMs, ChatGPT, in 83.33% of translation directions. As such, we have decided to use large machine translation models as a base for fine-tuning. We have decided to run experiments with M2M100 and NLLB-200 as base models, due to their performances with low-resource languages, as we have described in Section 3.3.3.

We fine-tuned the M2M100 (Fan et al., 2021) multilingual encoder-decoder, setting both the source

and target languages as Portuguese. We chose the M2M100_418M version of the M2M100[12] and the 600M version of the NLLB-200[13] due to temporal constraints.

5.2.2 Training

For training, we used the original sentences of our automatically curated set as source, and the respective versions rewritten by the RBM as target. We followed an 80-10-10 split for training, validation, and test sets.

We performed automated hyperparameter search for both models using Weights & Biases (Biewald, 2020). We ran a total of ten sweeps, exploring different combinations of values for learning rate and weight decay. The results are depicted in Figures 5.7 (relative to M2M100) and 5.8 (relative to NLLB-200).

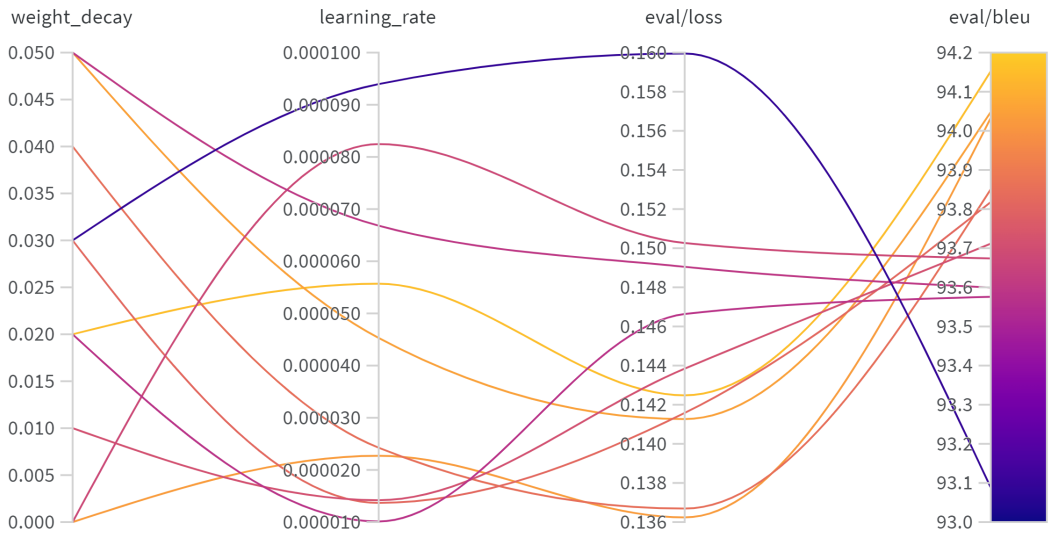

Figure 5.7: Hyperparameter search results for M2M100, provided by Weights & Biases (Biewald, 2020). The best configuration consists of a weight decay of 0.02 and a learning rate of 0.00005569.

The best parameters for the M2M100 (and therefore the ones we used during training) are defined as follows:

- **Learning Rate:** set at **0.00005569**
- **Weight Decay:** set at **0.02**
- **Batch Size:** set at **8**

[12] https://huggingface.co/facebook/m2m100_418M
[13] https://huggingface.co/facebook/nllb-200-distilled-600M

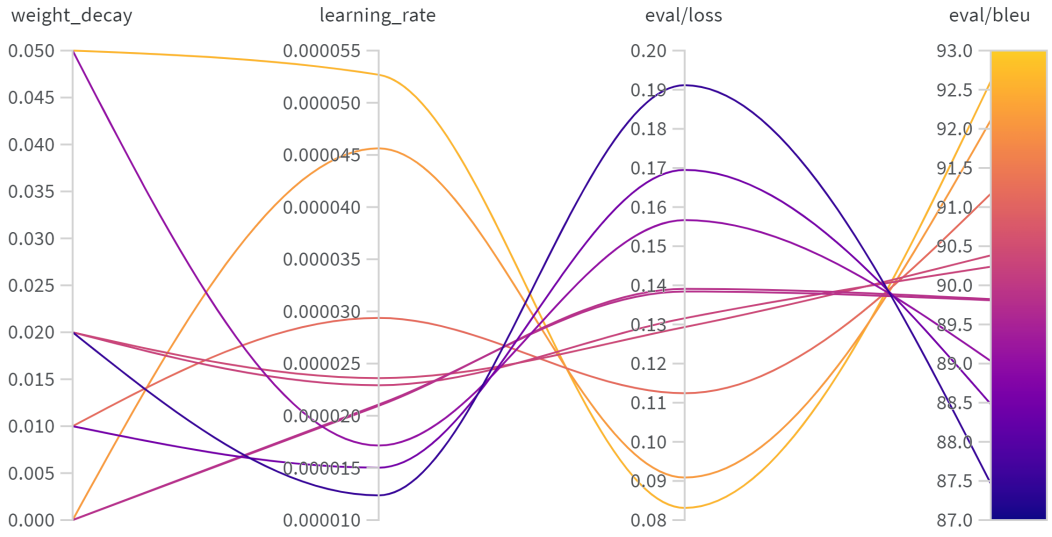

Figure 5.8: Hyperparameter search results for NLLB-200, provided by Weights & Biases (Biewald, 2020). The best configuration consists of a weight decay of 0.05 and a learning rate of 0.00005269.

- **Eval Batch Size:** set at **8**
- **Maximum Target Length:** set at **128**
- **Number of Epochs:** set at **5**

The best parameters for the NLLB-200 (and therefore the ones we used during training) are defined as follows:

- **Learning Rate:** set at **0.00005269**
- **Weight Decay:** set at **0.05**
- **Batch Size:** set at **8**
- **Eval Batch Size:** set at **8**
- **Maximum Target Length:** set at **128**
- **Number of Epochs:** set at **5**

For both models, the rest of the parameters are left as HuggingFace Seq2SeqTrainingArguments[14] defaults. We used the Seq2SeqTrainer[15] class to complete the fine-tuning procedure, using *sacrebleu* (Post, 2018) — a metric that handles downloading, processing, and tokenization, while producing the

[14] https://huggingface.co/docs/transformers/main_classes/trainer#transformers.Seq2SeqTrainingArguments
[15] https://huggingface.co/docs/transformers/main_classes/trainer#transformers.Seq2SeqTrainer

official BLEU WMT[16] scores — for our `compute_metrics` fuction. This function is used for evaluating the model after each epoch.

[16] https://machinetranslate.org/wmt Accessed on 18-09-2023

6

Experiments

Contents

We have previously introduced our datasets and models. In this chapter, we describe the chosen metrics, the evaluation setup, and respective results. Finally, we perform error analysis on the outputs of our rule-based model and our best neural model.

6.1 Metrics

Due to the novelty of the task of gender-neutral rewriting, we find that there is a lack of general consensus on useful metrics for evaluating these types of rewriter systems.

WER (Woodard and Nelson, 1982) Ratio of errors to total words in a text. It is used as metric for all gender-neutral rewriter systems described in Section 3.1. It can be computed as:

$$WER = (S + D + I)/N = (S + D + I)/(S + D + C)$$

where *S* is the number of substitutions, *D* is the number of deletions, *I* is the number of insertions, *C* is the number of correct words, and *N* is the number of words in the reference.

CER (Morris et al., 2004) Ratio of errors to total characters in a text. Since gender-neutral language in Romance languages often consists in a single character change (e.g. "João é bonit**o**." → "João é bonit**e**."), we believed this metric to be relevant for this task. It can be computed as:

$$CER = (S + D + I)/N = (S + D + I)/(S + D + C)$$

where *S* is the number of substitutions, *D* is the number of deletions, *I* is the number of insertions, *C* is the number of correct characters, and *N* is the number of characters in the reference *(N=S+D+C)*.

BLEU (Papineni et al., 2002) Evaluates the quality of a machine-translated text by comparing it with a quality reference translation. Can have a precision of 1-grams to 4-grams. In this work, we calculate BLEU-4 for easier comparison with the works described in Section 3.1, Since we consider the brevity penalty to be irrelevant for this task, we also calculate BLEU-1. It can be calculated as:

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where we use n-grams up to length *N*, positive weights w_n summing to one, and the geometric average of the modified n-gram precisions as p_n. *BP* corresponds to the brevity penanlty.

ROUGE-N (Lin, 2004) Is the n-gram recall between a source text and a set of references. We calculate ROUGE-1. ROUGE-N can be calculated as:

$$ROUGE\text{-}N = \frac{\sum_{S \in \{references\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{references\}} \sum_{gram_n \in S} Count(gram_n)}$$

where *n* stands for the length of the n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate text and a set of references.

6.2 Results

6.2.1 Neural Models Comparison

We compared the fine-tuned M2M-100 model and the NLLB-200 model (described in Section 5.2) to select our best neural model. The results are depicted in Table 6.1. The M2M-100 version outperforms or equals the performance of the NLLB-200 version in every metric, and was therefore selected as our best neural model. From this point on, this model is referred to simply as Neural Model (NM).

	M2M-100	NLLB-200
WER% ↓	**7.47**	7.95
CER% ↓	**1.95**	1.99
BLEU-4 ↑	**78.11**	77.226
BLEU-1 ↑	**92.21**	91.73
ROUGE-1 ↑	**93.41**	**93.41**

Table 6.1: Metrics for the test set (all text categories) of the fine-tuned versions of M2M-100 and NLLB-200. The best model for each category/metric pair is marked in bold.

6.2.2 Rule-based Model and Neural Model Comparison

We tested the RBM and the NM on our manually curated test set. We compared the results against a baseline metric (Base), which computes the metrics between the original examples and the gender-neutral versions. The results are detailed in Table 6.2.

The superior performance of the neural model on the dataset with simpler sentences suggests that the neural approach may be better suited for rewriting source sentences with a simpler syntax. While the average performance of the rule-based model tends to be slightly better compared to the performance of the neural model, the higher ROUGE score of the neural model suggests that this approach tends to rewrite fewer terms, since a higher ROUGE score corresponds to a higher rate of true positives.

		Literary	Journalistic	Dialogue	Social Media	Simple	Average
WER% ↓	Base	15.07	15.43	21.93	14.22	26.45	18.63
	RBM	7.89	**4.98**	**7.95**	**5.27**	7.02	**6.62**
	NM	**7.56**	8.03	8.19	7.35	**6.2**	7.47
CER% ↓	Base	3.90	2.86	7.20	3.75	7.36	5.01
	RBM	1.63	**1.03**	**2.26**	**1.23**	1.72	**1.57**
	NM	**1.55**	2.61	2.47	1.72	**1.39**	1.95
BLEU-4 ↑	Base	57.90	71.79	42.93	65.97	32.89	54.30
	RBM	76.28	**87.36**	**72.92**	**80.40**	74.37	**78.27**
	NM	**79.77**	83.91	71.60	78.53	**76.72**	78.106
BLEU-1 ↑	Base	81.00	85.13	74.53	83.72	71.23	79.12
	RBM	90.75	**94.32**	**90.75**	**93.25**	93.00	**92.414**
	NM	**92.11**	92.65	90.70	91.74	**93.83**	92.21
ROUGE-1 ↑	Base	83.15	90.80	79.15	86.63	74.41	82.93
	RBM	91.67	**95.00**	91.04	**93.77**	93.38	93.17
	NM	**92.95**	93.78	**92.35**	93.38	**94.59**	**93.41**

Table 6.2: Metrics for the manually curated test sets for each data category. The best model for each category/-metric pair is marked in bold.

6.3 Discussion

6.3.1 Overview

Rule-based Approach If the RBM produces a wrong output, the error can arise from one and/or two circumstances:

- **One of the preprocessing pipelines has produced an error:** for instance, the named entity recognition model may fail and wrongly tag a noun as a human referent, as in the case of the first example found in Table 6.4. Since we are using a Spanish NER model, this is a common mistake with names that may be unusual in Romance languages (such as “Bill”, found in the fourth example found in Table 6.4.) It might be worth to note that this problem is not specifically due to the NER model being focused on the Spanish language, but due to it being focused on a Romance language. These types of errors are not directly caused by the rules we have defined, and as such can only be mitigated by using different preprocessing tools.
- **The rewriter rules have failed:** The rules we have found to be most susceptible to errors are the ones regarding *adjectives*, as we have described in Section 5.1.4. Since Portuguese adjectives

are diverse in terms of gendered terminations, creating rules that encompass all adjectives in the language may be complex and time-consuming.

Neural Approach If the NM produces a wrong output, the error can arise from:

- **The model has learned a wrongly rewritten form of a certain term:** as is the case in the incorrect sentence depicted in Table 6.5. These types of errors may be mitigated by improving the quality of the rewriting rules.
- **The model is either ignoring terms that should be rewritten, or rewriting terms that should be ignored:** we have noticed that adjectives are more susceptible to these types of errors, most likely due to the model simply learning how to rewrite a certain term, but not capturing if the term is related (or not) to a human referent.

Similarly to Vanmassenhove et al. (2021), we have found that our neural model is able to generalize over the training data. For instance, the fifth sentence of Table 6.4 contains the expression *soprador de apito*[1]. Although the word *soprador* does not exist in the model training data, the model is able to rewrite the term as the respective gender-neutral form *sopradore*. However, this is not always the case: in the first sentence of Table 6.3, the neural model is not able to correctly rewrite the term *ricas*[2]. We hypothesize that this may be either due to the higher complexity of the respective gender neutral form (*ricas* $\rightarrow$ *riques*) or due to the lack of representation of female gender forms in data (both in our own datasets and the training data of our base neural model).

RBM	NM
Vocês são **riques**.	Vocês são **ricas**.
Elu está parade na parte mais **fria** do complexo.	Elu está parado na parte mais **frie** do complexo.
[...] foram convidadas **dues** profissionais **vindes** da RTP [...].	[...] foram convidadas **duas** profissionais **vindas** da RTP [...].
Em 1989 conseguiram eleger deputades para **o Parlamento Europeu**.	Em 1989 conseguiram eleger deputades para **Parlamento Europeu**.
Oi sou a sub **delegade**	Oi sou a sub **delegada**
Caso **Haddad** seja eleito, eu vou fazer diferente de alguns e vou pensar positivo [...]	Caso **o Haddad** seja eleito, eu vou fazer diferente de alguns e vou pensar positivo [...]

Table 6.3: Example sentences where the rule-based model performs better than the neural model.

[1] English: "whistle blower"
[2] English: "rich"

RBM	NM
Irene é **de** Peru.	Irene é **do** Peru.
Elu é a **maiora** pessoa que já viveu.	Elu é a **maior** pessoa que já viveu.
Eu dormi com a vítima número **dues**.	Eu dormi com a vítima número **dois**.
Minhe parente não me deixa sair com **o Bill**.	Minhe parente não me deixa sair com **Bill**.
[...] Tom trabalhava como **soprador** de apito [...].	[...] Tom trabalhava como **sopradore** de apito [...].
Boe dia criança !!!	**Bom** dia criança !!!

Table 6.4: Example sentences where the neural model performs better than the rule-based model.

	RBM	NM
Correct	**Ê presidente** se reunirá amanhã com **ês empresáries** mais importantes do país.	**Ê presidente** se reunirá amanhã com **ês empresáries** mais importantes do país.
Incorrect	Você é muito **gentile**.	Você é muito **gentile**.

Table 6.5: Example sentences where both models produce the same output (correctly or incorrecty).

6.3.2 Detailed Error Analysis

In order to perform a thorough manual error analysis, we have designed 17 error classes. Most of these classes are based on the word class that originated the error, as well as on the error itself: whether the term has not been neutralized when it should (NN) or has been wrongly neutralized (WN). Table 6.6 provides examples for each of the error labels.

Regarding the rule-based model outputs, verifying the exact point in the pipeline where the error has originated is a complex task. Since the neural model has a completely different architecture and nature, we have designed the error classes to be as architecture-independent as possible. This entails that two errors might belong to the same class but not have the same source. We follow with a brief explanation of the error labels, as well as their possible origin.

HR-NN A human referent was not neutralized when it should have been. In the context of the RBM, it is usually caused by the Human Referents Extractor module not tagging a term as a human referent. This may be either due to a POS-tagger error or a Wordnet error, such as the term not being present or the lexicographer file being incorrectly tagged.

HR-WN Either a human referent was neutralized (the term was wrongly tagged as a human referent), or the neutralization is incorrect. This may be due to a POS-tagger error, a problem related to Wordnet,

or the term (or, more concretely, its suffix) not being expected by the rules.

PN-WN A proper noun was not tagged as such, and therefore was wrongly neutralized. In the context of the RBM, this is caused by a NER model error.

NOUN-WN A non-human referent noun was wrongly tagged as one and therefore wrongly neutralized. In the context of the RBM, this is caused by a problem related to Wordnet.

ADJ-NN An adjective that is related to a human referent was not neutralized. In the context of the RBM, this is usually due to a failure of the rules regarding adjectives. In some rare cases, this may be due to a POS-tagger error.

ADJ-WN Either an adjective that is not related to a human referent was neutralized, or the neutralization is incorrect. In the first case, this is usually a consequence of a previous HR-WN error. In the second case, this is a rule failure (the term/suffix is not included in the rules).

DET-NN A determiner related to a human referent was not neutralized. In the context of the RBM, this may be due to a POS-tagger error or an error related to the NER model.

DET-WN A determiner that is not related to a human referent was wrongly neutralized. In the context of the RBM, this is usually due to a rule failure.

DET-NO A determiner that should have been omitted was not. In the context of the RBM, this is a consequence of a NER model error (see error HR-NN).

DET-WO A determiner was wrongly omitted. This is a consequence of a NER model error, where a noun is wrongly considered a proper noun (see error NOUN-WN).

PRON-NN A personal pronoun was not neutralized. This is either due to a POS-tagger error or a rule failure.

PRON-WN A personal pronoun was wrongly neutralized, or the neutralization is incorrect. This is usually due to all clitic pronouns being neutralized.

VERB-NN A verb that is related to a human referent was not neutralized. In the context of the RBM, this is usually due to a POS-tagger error.

VERB-WN A verb that is not related to a human referent was tagged as such, and therefore wrongly neutralized. In the context of the RBM, this is usually due to a rule failure or a POS-tagger error.

SPACE A whitespace was introduced or removed by the model or there was a typo in the dataset, causing an error. This error is almost exclusive to the NM.

OTHER A strange character was introduced by the model or there was a typo in the dataset, causing an error.

DUBIOUS The model output may also be considered gender-neutral and therefore correct, it simply is different from the annotated version.

The error counts for each model can be found in Table 6.7. The error counts for the errors consisting of wrong neutralizations (the error labels that contain -WN) are higher in the RBM. This confirms that the NM tends to rewrite fewer terms, as we have mentioned in Section 6.2. However, the counts for the SPACE and OTHER errors are higher in the NM outputs, suggesting that model hallucinations and insertions/removal of spaces can have an impact on the performance of our NM. This is relevant for the CER metric. WER, BLEU, and ROUGE do not take into account the removal/insertion of spaces.

The three most common types of errors, taking into account both models, are HR-NN, ADJ-NN, and PRON-NN. This provides us with some insight into possible directions for future corrections of the models, especially the RBM. In the case of HR-NN errors, we are somehow limited to the capacity of our chosen Wordnet, since most HR-NN (in the context of the RBM) arise from the extractor module not tagging a certain term as a human referent. However, we believe many ADJ-NN and PRON-NN errors may be mitigated by altering and adding new rewriter rules. Due to time constraints, the analysis of exactly which rules may be altered or added is left as future work.

Error Class	Example	Model Output
HR-NN	Eu sei que Tom é **ume estrangeire**.	Eu sei que Tom é **um estrangeiro**.
HR-WN	Ê **guitarrista** lutadore?	Ê **guitarriste** lutadore?
PN-WN	**Alice dos Santos** ia com o comite diretor ao jardim público [...]	**Alice des Santos** ia com o comite diretor ao jardim público [...]
NOUN-WN	Eu estou encantade com o desempenho **do computador**.	Eu estou encantade com o desempenho **de computadore**.
ADJ-NN	Vocês são **riques**.	Vocês são **ricas**.
ADJ-WN	Você é muito **gentil**.	Você é muito **gentile**.
DET-NN	Eu sou **ê** únique que sobreviveu ao acidente.	Eu sou **o** únique que sobreviveu ao acidente.
DET-WN	E assim cada ume de nós é **um** pouco criadore [...]	E assim cada ume de nós é **ume** pouco criadore [...]
DET-NO	Eu ligo para **Tom** quase todos os dias.	Eu ligo para **o Tom** quase todos os dias.
DET-WO	Em 1989 conseguiram eleger deputades para **o Parlamento Europeu**.	Em 1989 conseguiram eleger deputades para **Parlamento Europeu**.
PRON-NN	Tom amou Mary e Mary **ê** amou.	Tom amou Mary e Mary **o** amou.
PRON-WN	Tudo o que ê senhore fizer, faça-**o** prontamente.	Tudo o que ê senhore fizer, faça-**e** prontamente.
VERB-NN	Elu foi **pegue** de surpresa.	Elu foi **pega** de surpresa.
VERB-WN	Eu **preciso** tentar encontrá-le.	Eu **precise** tentar encontrá-le.
SPACE	Ê presidente fará um discurso no Dia dos Mortos**.....** no povoado de Culiácan.	Ê presidente fará um discurso no Dia dos Mortos**.....no** povoado de Culiácan.
OTHER	**Antropólogue**.	**Antropólogo**.
DUBIOUS	**A criança** é **bonita**.	**Ê menininhe** é **bonite**.

Table 6.6: Error labels and respective examples. The "Example column" contains manually annotated gender-neutral sentences. The "Model Output" column contains incorrect outputs from one of our models. The differences between the sentences (which correspond to the errors) are tagged in bold.

Error Class	RBM	NM
HR-NN	73	60
HR-WN	18	19
PN-WN	4	1
NOUN-WN	22	19
ADJ-NN	24	35
ADJ-WN	24	17
DET-NN	50	49
DET-WN	8	9
DET-NO	8	8
DET-WO	5	5
PRON-NN	10	13
PRON-WN	2	0
VERB-NN	5	1
VERB-WN	10	3
SPACE	1	20
OTHER	2	9
DUBIOUS	4	4
Total	270	272

Table 6.7: Error classes and respective error counts, regarding the outputs of the RBM and the NM.

7

Conclusion

Contents

7.1 Contributions

This work is a first effort towards creating NLP resources and models that contain and are able to correctly process gender-neutral Portuguese. We present the first Portuguese dataset explicitly containing gender-neutral language and neopronouns, along with a rule-based and a neural gender-neutral rewriters. Additionally, we provide a manually annotated collection of 500 original sentences and a respective gender-neutral version. One entry of our dataset consists of a binary-gendered sentence with the respective gender-neutral version provided by the RBM. An automatically generated set of this parallel data was used for training the NM. We provide the first benchmarks of the gender-neutral rewriting task for the Portuguese language.

7.2 Limitations and Future Work

Due to the novelty of the gender-neutral rewriting task, as well as the constant advancements in the area and new approaches to the task, there are many possible new avenues for future work.

- The usage of gender-neutral language in Portuguese-speaking communities is a diverse and ever-changing linguistic phenomenon. While we present some of the third neutral gender pronouns found in literature in Section 2.3, our models only process the neopronoun *elu* and follow rewriting rules that are not universally agreed upon. Therefore, one priority of future work should be the inclusion of other neopronouns.
- Our rule-based model suffers from low scalability to long text. Handmade rules often fail to correctly rewrite long sentences due to their more complex and unpredictable structure. Furthermore, the current version of our rule-based model could be greatly improved by altering and adding new rules.
- While using pre-trained large multilingual translation models may be an option for developing gender-neutral rewriters for lower-resource languages, this method is dependent on the existence of said models for the target language. Languages with very low resources are often not represented in such models.
- Our rule-based model only functions at a sentence-level. In order to create dependencies between sentences and linking human referents and dependent terms between sentences, our model should be enriched with a coreference resolution module. This entails annotating gender-inclusive data (for example, our own datasets) for the task of coreference resolution.
- Although the neural model can generalize over the seen data, we hypothesize that it fails to internalize the context of sentences and whether gendered terms refer to humans or objects. In future

work, we hope to bypass this issue either by training a larger model with more quality data, and/or developing a hybrid rule-based/neural model. Another possible approach is data augmentation using LLM prompting. In a first attempt to replicate this approach, we have prompted ChatGPT[1], the sibling model to InstructGPT (Ouyang et al., 2022), to generate Portuguese gender-neutral sentences. Results are depicted in Appendix C. Although a promising approach, we hypothesize that the generated sentences do not have the necessary consistency and variety to create a quality dataset. Furthermore, at the time of writing, LLMs are still performing poorly when tasked with predicting the correct forms of neopronouns (Hossain et al., 2023). We expect that, in the future, with the advancement of these types of models and optimized prompting, we are able to generate quality gender-neutral examples that allow us to create larger inclusive datasets.

[1] chat.openai.com

Bibliography

L. Biewald, “Experiment tracking with weights and biases,” 2020, software available from wandb.com. [Online]. Available: https://www.wandb.com/

Merriam-Webster, “Merriam-Webster Dictionary,” https://www.merriam-webster.com/, 2023, accessed: 2022-10-13.

L. Hekanaho, “Generic and nonbinary pronouns: Usage, acceptability and attitudes,” *Neuphilologische Mitteilungen*, vol. 121, no. 2, pp. 498–509, 2020.

S. Gal, “Between speech and silence: The problematics of research on language and gender,” *IPrA Papers in Pragmatics*, vol. 3, no. 1, pp. 1–38, 1989.

——, “Peasant men can’t get wives: Language change and sex roles in a bilingual community,” *Language in society*, vol. 7, no. 1, pp. 1–16, 1978.

L. R. R. Pinheiro, “Linguagem neutra: a reestruturação do gênero no português brasileiro frente às mudanças sociais,” https://bdm.unb.br/handle/10483/28202, 2020, accessed: 18-09-2023.

M. J. R. Miranda, “Português para todes?: um diálogo entre a análise de discurso crítica e a sociolinguística sobre linguagem não binária,” https://bdm.unb.br/handle/10483/28244, 2020, accessed: 01-12-2022.

P. Zhou, W. Shi, J. Zhao, K.-H. Huang, M. Chen, R. Cotterell, and K.-W. Chang, “Examining gender bias in languages with grammatical gender,” 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)*, 2019, pp. 5276–5284.

E. Vanmassenhove, C. Emmery, and D. Shterionov, “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*, 2021, pp. 8940–8948.

T. Sun, K. Webster, A. Shah, W. Y. Wang, and M. Johnson, “They, them, theirs: Rewriting with gender-neutral english,” *arXiv preprint arXiv:2102.06788*, 2021.

K. Slemp, “Latino, latina, latin@, latine, and latinx: gender inclusive oral expression in spanish,” Ph.D. dissertation, The University of Western Ontario (Canada), 2020.

J. M. Sevelius, “Gender affirmation: A framework for conceptualizing risk behavior among transgender women of color,” *Sex roles*, vol. 68, no. 11, pp. 675–689, 2013.

J. M. Sevelius, D. Chakravarty, S. E. Dilworth, G. Rebchook, and T. B. Neilands, “Gender affirmation through correct pronoun usage: development and validation of the transgender women’s importance of pronouns (tw-ip) scale,” *International journal of environmental research and public health*, vol. 17, no. 24, p. 9525, 2020.

F. Hamidi, M. K. Scheuerman, and S. M. Branham, “Gender recognition or gender reductionism? the social implications of embedded gender recognition systems,” in *Proceedings of the 2018 chi conference on human factors in computing systems*, 2018, pp. 1–13.

J. A. Sniezek and C. H. Jazwinski, “Gender bias in english: In search of fair language,” *Journal of Applied Social Psychology*, vol. 16, no. 7, pp. 642–662, 1986.

C. Kaufmann and G. Bohner, “Masculine generics and gender-aware alternatives in spanish,” *IZGOnZeit. Onlinezeitschrift des Interdisziplinären Zentrums für Geschlechterforschung (IZG)*, pp. 8–17, 2014.

D. Jurafsky and J. H. Martin, “Speech and language processing (draft),” *preparation [cited 2022 November 3] Available from: https://web. stanford. edu/˜ jurafsky/slp3*, 2021.

G. G. Corbett *et al.*, *Gender*. Cambridge University Press, 1991.

C. F. Hockett, *A course in modern linguistics.* Macmillan, 1967.

G. Lakoff, *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press, 2008.

M. Auxland, “Para todes: A case study on portuguese and gender-neutrality,” *Journal of Languages, Texts and Society*, vol. 4, pp. 1–23, 2020.

T. Konishi, “The semantics of grammatical gender: A cross-cultural study,” *Journal of psycholinguistic research*, vol. 22, no. 5, pp. 519–534, 1993.

W. Phillips and L. Boroditsky, “Can quirks of grammar affect the way you think? grammatical gender and object concepts,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2003.

S. Stryker, *Transgender history: The roots of today’s revolution*. Hachette UK, 2017.

J. Tyroler, “Transmission-interview with Leslie Feinberg,” *Camp Kansas City*, 2006.

J. Hokenson, “The pronouns of Gomorrha: A lesbian prose tradition,” *Frontiers: A Journal of Women Studies*, pp. 62–69, 1988.

B. A. Rogers, “Drag as a resource: Trans* and nonbinary individuals in the southeastern united states,” *Gender & Society*, vol. 32, no. 6, pp. 889–910, 2018.

M. Gustafsson Sendén, E. Renström, and A. Lindqvist, “Pronouns beyond the binary: The change of attitudes and use over time,” *Gender & Society*, vol. 35, no. 4, pp. 588–615, 2021.

S. Sczesny, M. Formanowicz, and F. Moser, “Can gender-fair language reduce gender stereotyping and discrimination?” *Frontiers in psychology*, p. 25, 2016.

D. Stahlberg, S. Sczesny, and F. Braun, “Name your favorite musician: Effects of masculine generics and of their alternatives in german,” *Journal of Language and Social Psychology*, vol. 20, no. 4, pp. 464–469, 2001.

L. C. Hord, “Bucking the linguistic binary: Gender neutral language in english, swedish, french, and german,” *Western Papers in Linguistics*, vol. 3, no. 1, 2016.

P. Berlucci and A. Zanella, “Manifesto ILE,” https://diversitybbox.com/manifesto-ile-para-uma-comunicacao-radicalmente-inclusiva/, accessed: 2022-10-28.

G. Caê, *Manual para o uso da linguagem neutra em Língua Portuguesa*, 2020.

D. Santos and V. Marques, “Guia prático para um português inclusivo,” http://queerist.tecnico.ulisboa.pt/projetos/linguagem, 2021, accessed: 2022-10-28.

P. Valente, “Sistema elu, linguagem neutra em género,” https://dezanove.pt/sistema-elu-linguagem-neutra-em-genero-1317469, 2020, accessed: 2022-10-30.

F. Fernández, “Transliteration,” *Gender outlaws: The next generation*, pp. 128–133, 2010.

H. Abe, *Queer Japanese: Gender and sexual identities through linguistic practices*. Springer, 2010.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

T. Diesner-Mayer and N. Seidel, “Supporting gender-neutral writing in german,” in *Proceedings of Mensch Und Computer 2022*, ser. MuC ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 509–512. [Online]. Available: https://doi.org/10.1145/3543758.3547566

M. Carl, S. Garnier, J. Haller, A. Altmayer, and B. Miemietz, “Controlling gender equality with shallow nlp techniques,” in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 2004, pp. 820–826.

C. Amrhein, F. Schottmann, R. Sennrich, and S. Läubli, “Exploiting biased models to de-bias text: A gender-fair rewriting model,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 4486–4506. [Online]. Available: https://aclanthology.org/2023.acl-long.246

V. Bellandi and S. Siccardi, “Gender discriminatory language identification with an hybrid algorithm based on syntactic rules and machine learning,” in *The 30th Italian Symposium on Advanced Database Systems*, 2022.

B. Alhafni, N. Habash, and H. Bouamor, “User-centric gender rewriting,” Association for Computational Linguistics (ACL), pp. 618–631, 2022.

B. Alhafni, O. Obeid, and N. Habash, “The user-aware arabic gender rewriter,” in *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*. Tampere, Finland: European Association for Machine Translation, Jun. 2023, pp. 3–11. [Online]. Available: https://aclanthology.org/2023.gitt-1.1

S. Dev, M. Monajatipoor, A. Ovalle, A. Subramonian, J. Phillips, and K.-W. Chang, “Harms of gender exclusivity and challenges in non-binary representation in language technologies,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 1968–1994.

R. Munro and A. C. Morrison, “Detecting independent pronoun bias with partially-synthetic data generation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2011–2017.

S. Brandl, R. Cui, and A. Søgaard, “How conservative are language models? adapting to the introduction of gender-neutral pronouns,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 3624–3630.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451.

“Neuralcoref 4.0,” https://github.com/huggingface/neuralcoref, accessed: 2022-11-03.

R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme, “Gender bias in coreference resolution,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 8–14.

A. Lauscher, A. Crowley, D. Hovy *et al.*, "Welcome to the modern world of pronouns: identity-inclusive natural language processing beyond gender," in *Proceedings of the 29th International Conference on Computational Linguistics COLING 2022*. (seleziona...), 2022.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

D. Saunders, R. Sallis, and B. Byrne, "Neural machine translation doesn't translate gender coreference right unless you make it," in *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 2020, pp. 35–43.

G. Stanovsky, N. A. Smith, and L. Zettlemoyer, "Evaluating gender bias in machine translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1679–1684.

E. Vanmassenhove, C. Hardmeier, and A. Way, "Getting gender right in neural machine translation," in *2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018, pp. 3003–3008.

W. I. Cho, J. W. Kim, S. M. Kim, and N. S. Kim, "On measuring gender bias in translation of gender-neutral pronouns," in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 2019, pp. 173–181.

Y. T. Cao and H. Daumé, "Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle," *Computational Linguistics*, vol. 47, no. 3, pp. 615–661, 2021.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 15–20.

G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

V. de Paiva, A. Rademaker, and G. de Melo, "Openwordnet-pt: An open Brazilian Wordnet for reasoning," in *Proceedings of COLING 2012: Demonstration Papers*. Mumbai, India: The COLING 2012 Organizing Committee, Dec. 2012, pp. 353–360, published also as Techreport http://hdl.handle.net/10438/10274. [Online]. Available: http://www.aclweb.org/anthology/C12-3044

P. University, "Princeton university "about wordnet."," https://wordnet.princeton.edu/, 2010, accessed: 2022-11-03.

H. Gonçalo Oliveira and P. Gomes, “Eco and onto. pt: a flexible approach for creating a portuguese wordnet automatically,” *Language resources and evaluation*, vol. 48, no. 2, pp. 373–393, 2014.

P. Marrafa, “Portuguese wordnet: general architecture and internal semantic relations,” *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, vol. 18, pp. 131–146, 2002.

P. Marrafa, R. Amaro, and S. Mendes, “Wordnet. pt global–extending wordnet. pt to portuguese varieties,” in *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, 2011, pp. 70–74.

S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O’Reilly Media, Inc., 2009.

A. Rademaker, F. Chalub, L. Real, C. Freitas, E. Bick, and V. de Paiva, “Universal dependencies for portuguese,” in *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, Pisa, Italy, September 2017, pp. 197–206. [Online]. Available: http://aclweb.org/anthology/W17-6523

P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza: A python natural language processing toolkit for many human languages,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 101–108.

N. Mamede, J. Baptista, C. Diniz, and V. Cabarrão, “String: An hybrid statistical and rule-based natural language processing chain for portuguese,” in *Computational Processing of the Portuguese Language, Proceedings of the 10th International Conference, PROPOR*, 2012, pp. 17–20.

P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, “The state and fate of linguistic diversity and inclusion in the nlp world,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6282–6293.

M. Przystupa and M. Abdul-Mageed, “Neural machine translation of low-resource and similar languages with backtranslation,” in *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, 2019, pp. 224–235.

A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary *et al.*, “Beyond english-centric multilingual machine translation,” *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 4839–4886, 2021.

M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard *et al.*, “No language left behind: Scaling human-centered machine translation,” *arXiv preprint arXiv:2207.04672*, 2022.

Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

D. Santos, R. Willrich, M. Langfeldt, R. G. de Moraes, C. Mota, E. Pires, R. Schumacher, and P. S. Pereira, "Identifying literary characters in portuguese: Challenges of an international shared task," in *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*. Springer, 2022, pp. 413–419.

D. Ameixa and L. Coheur, "From subtitles to human interactions: introducing the subtle corpus," INESC-ID, Tech. Rep., 2013.

D. Ameixa, L. Coheur, P. Fialho, and P. Quaresma, "Luke, i am your father: dealing with out-of-domain requests by using movies subtitles," in *Intelligent Virtual Agents: 14th International Conference, IVA 2014, Boston, MA, USA, August 27-29, 2014. Proceedings 14*. Springer, 2014, pp. 13–21.

J. Tiedemann, "The tatoeba translation challenge – realistic data sets for low resource and multilingual MT," in *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1174–1182. [Online]. Available: https://aclanthology.org/2020.wmt-1.139

A. Morris, V. Maier, and P. Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition." 01 2004.

R. W. Langacker, *Language and its structure*. Harcourt Brace Jovanovich New York, NY, 1973.

A. Branco, J. Silva, F. Costa, and S. Castro, "Cintil treebank handbook: Design options for the representation of syntactic constituency," Technical Report TR-2011-02. Available at: http://docs. di. fc. ul. pt, Tech. Rep., 2011.

M. Gonçalves, L. Coheur, J. Baptista, and A. Mineiro, "Avaliação de recursos computacionais para o português," *Linguamática*, vol. 12, no. 2, pp. 51–68, 2021.

M. W. Goodman and F. Bond, "Intrinsically interlingual: The wn python library for wordnets," *South African Centre for Digital Language Resources (SADiLaR) Potchefstroom, South Africa*, p. 100, 2021.

E. Loper and S. Bird, "Nltk: The natural language toolkit," in *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 214–217. [Online]. Available: https://aclanthology.org/P04-3031

K. K. Zethsen, "Intralingual translation: An attempt at description," *Meta*, vol. 54, no. 4, pp. 795–812, 2009.

W. Zhu, H. Liu, Q. Dong, J. Xu, L. Kong, J. Chen, L. Li, and S. Huang, “Multilingual machine translation with large language models: Empirical results and analysis,” *arXiv preprint arXiv:2304.04675*, 2023.

M. Post, “A call for clarity in reporting BLEU scores,” in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: https://www.aclweb.org/anthology/W18-6319

J. Woodard and J. Nelson, “An information theoretic measure of speech recognition performance,” 1982.

K. Papineni, S. Roukos, T. Ward, and W. jing Zhu, “Bleu: a method for automatic evaluation of machine translation,” 2002, pp. 311–318.

C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://www.aclweb.org/anthology/W04-1013

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, “Training language models to follow instructions with human feedback,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

T. Hossain, S. Dev, and S. Singh, “Misgendered: Limits of large language models in understanding pronouns,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 5352–5367. [Online]. Available: https://aclanthology.org/2023.acl-long.293

A

Gender-neutral Grammar Details and Examples

A.1 Examples from Users of Portuguese Gender-neutral Language

Algum amigue pelo litoral de cabedelo para um rolê na praia?
Minha amigue tirou a runa das bruxas para mim e eu estou impactada
Todas e todes recebendo tratamento igualitário. . .
Procuro um namorade que me dê carinho e atenção

Table A.1: Example excerpts retrieved from Twitter in 20/02/2023. We slightly modified the examples to lower searchability and increase the privacy of the authors.

A.2 Gender-neutral Expressions

Original Expression	Gender-neutral Expression
homem/mulher	pessoas
rapaz/rapariga	jovem
menino/menina	criança
pai/mãe	parente
aluno/aluna	estudante
professor/professora	docente
esposo/esposa	cônjuge
rei/rainha	monarca

Table A.2: Binary-gendered expressions and respective gender-neutral alternative expressions.

A.3 Example for a Third Neutral Gender Grammar

	Personal Pronouns		
Masculine	ele(s)	o(s)	lo(s)
Feminine	ela(s)	a(s)	la(s)
Neutral	elu(s)	ê(s)	le(s)

	Possessive Pronouns				
Masculine	meu(s)	teu(s)	seu(s)	nosso(s)	vosso(s)
Feminine	minha(s)	tua(s)	sua(s)	nossa(s)	vossa(s)
Neutral	minhe(s)	tue(s)	sue(s)	nosse(s)	vosse(s)

	Demonstrative Pronouns					
Masculine	este(s)	esse(s)	aquele(s)	mesmo(s)	outro(s)	tanto(s)
Feminine	esta(s)	essa(s)	aquela(s)	mesma(s)	outra(s)	tanta(s)
Neutral	estu(s)	essu(s)	aquelu(s)	mesme(s)	outre(s)	tante(s)

	Relative and Interrogative Pronouns	
Masculine	cujo(s)	quanto(s)
Feminine	cuja(s)	quanta(s)
Neutral	cuje(s)	quante(s)

	Undefined Pronouns								
Masculine	muito	pouco	tanto	todo	nenhum	algum	certo	outro	ambos
Feminine	muita	pouca	tanta	toda	nenhuma	alguma	certa	outra	ambas
Neutral	muite	pouque	tante	tode	nenhume	algume	certe	outre	ambes

	Articles	
Masculine	o(s)	um(ns)
Feminine	a(s)	uma(s)
Neutral	ê(s)	ume(s)

	Prepositions			
Masculine	pelo	do	no	ao
Feminine	pela	da	na	à
Neutral	pele	de	ne	ae

Nouns and Adjectives			
Termination\Gender	Masculine	Feminine	Neutral
-o/-a/-e	filho	filha	filhe
-co/-ca/-que	técnico	técnica	técnique
-go/-ga/-gue	amigo	amiga	amigue
-ão/-ã/-ãe	irmão	irmã	irmãe
-ão/-ona/-one	chorão	chorona	chorone
-ão/-oa/-oe	patrão	patroa	patroe
-r/-ra/-re	professor	professora	professore
-tor/-triz/-tore	ator	atriz	atore
-e/-a/-e	governante	governanta	governante
-ês/esa/ese	burguês	burguesa	burguese
-z/za/ze	juiz	juíza	juíze
-l/-la/-le	bacharel	bacharela	bacharele
-u/-ua/-ue	nu	nua	nue
-eu/-eia/-eie	ateu	ateia	ateie
-ois/-uas/-ues	dois	duas	dues

B

Annotation Guidelines

These annotation guidelines contain rewriting rules for every gendered word class. Although the original guide was written in Portuguese, here we also present an English version of the text.

B.1 Introdução / Introduction

PT Este guia contém as diretrizes para a tarefa de reescrita de texto segundo o sistema "elu" de linguagem género-neutra. A compilação destas diretrizes advém da necessiadade de criação de dados de linguagem género-neutra para o Português, para efeitos de treinar modelos de linguagem capazes de processar diferentes pronomes e linguagem género-neutra. Com a criação destas diretrizes, não afirmamos que o uso do pronome "elu" seja de algum modo superior ao uso de outros pronomes género-neutros usados pela comunidade lusófona, bem como não afirmamos que a sintaxe usada para criar concordância com pronomes género-neutros é de algum modo mais "correta" do que outras. Como referência, usamos os guias elaborados pelo QueerIST[1] Santos and Marques (2021) e Caê Alemeida Caê (2020). Utilizamos as regras referentes ao sistema "elu" por simplicidade, por ser o pronome neutro

[1] http://queerist.tecnico.ulisboa.pt/

mais reportado na comunidade lusófona, e também por conter menos carateres com acentuação nas respetivas formas neutras.

EN This document contains guidelines for the gender-neutral rewriting task according to the "elu" gender-neutral language system. These guidelines originates from the need to create gender-neutral language data for Portuguese, in order to train language models capable of processing different pronouns and gender-neutral language. With the creation of these guidelines, we do not claim that the use of the pronoun "elu" is in any way superior to the use of other gender-neutral pronouns used by the Portuguese-speaking community, nor do we claim that the syntax used to create agreement with gender-neutral pronouns is in any way more "correct" than others. As a reference, we used the guides prepared by QueerIST[2] Santos and Marques (2021) and Caê Alemeida Caê (2020). We used the rules referring to the "elu" system due to simplicity, as it is the most reported neutral pronoun in the Portuguese-speaking community, and because respective neutral forms tend to contain fewer accented characters.

B.2 Classes de Palavras e Respetivas Regras / Word Classes and Associated Rules

PT De uma forma geral, qualquer instância de um termo masculino ou feminino que se refere a uma pessoa (ou que deve ter concordâcia de género com um termo relacionado) deve ser neutralizado. Da mesma forma que a terminação masculina é **-o** e a terminação feminina é **-a**, a terminação neutra é **-e**, exceto para palavras em que a terminação masculina é também **-e**. Nesses casos, a terminação neutra é **-u**, de forma a não se confundir com a forma masculina.

EN Generally, any instance of a masculine or feminine term that refers to a person (or is gender-concordant with a related term) should be neutralized. Just as the masculine suffix is **-o** and the feminine suffix is **-a**, the neutral suffix is **-e**. Words where the masculine suffix is also **-e** are an exception: in those cases, the neutral suffix is **-u**, so as not to be confused with the masculine form.

B.2.1 Pronomes / Pronouns

PT Assumimos que todos os pronomes **pessoais** se referem a pessoas (**independentemente do contexto da frase**) e como tal devem ser neutralizados. Não assumimos que as restantes subclasses de pronomes se referem a pessoas, logo só as neutralizamos se estiverem relacionadas com um termo que se refere a uma pessoa.

[2] http://queerist.tecnico.ulisboa.pt/

EN We assume that all **personal** pronouns refer to people (**regardless of the context of the sentence**) and as such must be neutralized. We do not assume that the other pronoun subclasses refer to people, and therefore we only neutralize them if they are related to a term that refers to a person.

B.2.1.A Pronomes Pessoais / Personal Pronouns

ele(s)/ela(s) → elu(s)

- Ele é adorável.[3] → Elu é adorável.

le(s)/la(s) → le(s)

- Hoje vou vê-lo.[4] → Hoje vou vê-le.

mo(s)/ma(s) → me(s)

- Ele lembra-mo.[5] → Elu lembra-me.

o(s)/a(s) → e(s)

- Deixa-o.[6] → Deixa-e.

B.2.1.B Pronomes Possessivos / Possessive Pronouns

meu(s)/minha(s) → minhe(s)

- Ele é meu colega.[7] → Elu é minhe colega.

teu(s)/tua(s) → tue(s)

- O Baltasar é teu amigo.[8] → Baltasar é tue amigue.

seu(s)/sua(s) → sue(s)

- O João é o seu pai.[9] → João é ê sue pai.

nosso(s)/nossos(s) → nosses(s)

- Eles são os nossos amigos.[10] → Elus são ês nosses amigues.

vosso(s)/vossa(s) → vosse(s)

[3] English: "He is adorable."
[4] English: "I'm going to see him today."
[5] English: "He reminds me of him."
[6] English: "Leave him."
[7] English: "He is my colleague."
[8] English: "Baltasar is your friend."
[9] English: "John is his father."
[10] English: "They are our friends."

- Eles são os vossos pais.[11] → Elus são ês vosses parentes.

B.2.1.C Pronomes Demonstrativos / Demonstrative Pronouns

este(s)/esta(s) → estu(s)

- Este é o meu namorado.[12] → Estu é ê minhe namorade.

esse(s)/essa(s) → essu(s)

- Essa é a tua irmã.[13] → Essu é ê tue irmãe.

aquele(s)/aquela(s) → aquelu(s)

- Aquele moço é bonito.[14] → Aquelu moce é bonite .

mesmo(s)/mesma(s) → mesme(s)

- Elas têm os mesmos amigos.[15] → Elus têm ês mesmes amigues.

outro(s)/outra(s) → outre(s)

- Eu quero outro presidente.[16] → Eu quero outre presidente.

tanto(s)/tanta(s) → tante(s)

- Estão a passar tantos corredores![17] → Estão a passar tantes corredores!

B.2.1.D Pronomes Indefinidos / Undefined Pronouns

muito(s)/muita(s) → muite(s)

pouco(s)/pouca(s) → pouque(s)

tanto(s)/tanta(s) → tante(s)

todo(s)/toda(s) → tode(s)

nenhum(s)/nenhuma(s) → nenhume(s)

algum(s)/alguma(s) → algume(s)

[11] English: “They are your parents.”
[12] English: “This is my boyfriend.”
[13] English: “That is your sister.”
[14] English: “That guy is pretty.”
[15] English: “They have the same friends.”
[16] English: “I want another president.”
[17] English: “There are so many runners passing by!”

certo(s)/certa(s) → **certe(s)**

outro(s)/outra(s) → **outre(s)**

ambos/ambas → **ambes**

B.2.2 Determinantes / Determiners

PT Se precedidos por um nome próprio, os determinantes artigos definidos (o, a, os, as) devem ser omitidos. Se precedidos por um nome comum, não devem ser omitidos. Os determinantes artigos indefinidos (um, uma, uns, umas) nunca são omitidos e, se referentes a uma pessoa, devem ser reescritos com a respetiva forma neutra.

EN If preceded by a proper noun, the definite article determiners (o, a, os, as) must be omitted. If preceded by a common noun, they must not be omitted. The indefinite article determiners (um, uma, uns, umas) are never omitted and, if they refer to a person, must be rewritten according to the respective neutral form.

o(s)/a(s) → **ê(s)/omitido**

- O Sérgio é o amigo da Mariana.[18] → Sérgio é ê amigue de Mariana.
- O João é simpático.[19] → João é simpátique.

um/uma(s) → **ume(s)**

- O Miguel é um escritor.[20] → Miguel é ume escritore.

B.2.3 Contrações com Proposições

PT Os casos específicos das contrações **pelo**/**pela** e **ao**/**à** possuem uma particularidade: no caso de serem precedidas por um nome próprio, deve-se optar pela 1ª forma neutra apresentada (**por** ou **a**). No caso de serem precedidas por outro tipo de nome, deve-se optar pela 2ª forma neutra apresentada (**pele** ou **ae**).

EN For the particular cases of the contractions **pelo**/**pela** and **ao**/**à**: in case they are preceded by a proper noun, one should opt for the 1st presented neutral form (**por** or **a**). If they are preceded by another type of noun, one must choose the 2nd neutral form presented (**pele** or **ae**).

[18] English: “Sérgio is Mariana’s friend.”
[19] English: “João is nice.”
[20] English: “Miguel is a writer.”

pelo(s)/pela(s) → por/pele(s)

- A Mariana fez tudo pela Lúcia.[21] → Mariana fez tudo por Lúcia.
- A Mariana fez tudo pelas amigas.[22] → Mariana fez tudo peles amigues.

ao(s)/à(s) → a(s)/ae(s)

- A Teresa deu o livro ao Pedro.[23] → Teresa deu o livro a Pedro.
- A Teresa deu o livro ao amigo.[24] → Teresa deu o livro ae amigue.

PT Os restantes casos são simples.

EN The other cases are simple.

do(s)/da(s) → de(s)

- O livro é da Teresa.[25] → O livro é de Teresa.

no(s)/na(s) → ne(s)

- A Teresa pôs o chapéu no amigo.[26] → Teresa pôs o chapéu ne amigue.

num/numas(s) → nume(s)

- Ele tropeçou numa menina.[27] → Elu tropeçou nume menine.

dele(s)/dela(s) → delu(s)

- O livro é dela.[28] → O livro é delu.

desse(s)/dessa(s) → dessu(s)

- O livro é desses rapazes.[29] → O livro é dessus jovens.

B.2.4 Nomes e Adjetivos / Nouns and Adjectives

PT Genericamente, palavras masculinas terminadas em -o ou palavras femininas terminadas em -a passam a ter uma terminação em -e. A formação geral de termos com terminação em -e segue as seguintes regras:

[21] English: “Mariana did everything for Lúcia.”
[22] English: “Mariana did everything for her friends.”
[23] English: “Teresa gave Pedro the book.”
[24] English: “Teresa gave the book to her friend.”
[25] English: “The book belongs to Teresa.”
[26] English: “Teresa put the hat on her friend.”
[27] English: “He tripped on a girl.”
[28] English: “The book is hers.”
[29] English: “The book belongs to those boys.”

EN Generally, masculine words ending in -o or feminine words ending in -a are rewritten as to terminate in -e. The general formation of terms ending in -e adheres to the following rules:

1. Se a forma feminina da palavra termina em -a, retira-se o -a e acrescenta-se um -e. / If the feminine form ends in -a, we remove the -a and add an -e.
2. Se uma das formas termina em -e, essa passa a ser a forma neutra. / If one of the forms ends in -e, that becomes the neutral form as well.

PT Seguem-se exemplos de algumas terminações.

EN We follow with examples for a few different suffixes.

-ão/-ã → -ãe

- O João é meu irmão.[30] → João é minhe irmãe.

-ão/-ona → -one

- A Lúcia é uma valentona.[31] → Lúcia é ume valentone.

-tor/-triz → -tore

- A Mariana é atriz.[32] → Mariana é atore.

-om/-oa → -oe

- A Dani é boa cozinheira.[33] → Dani é boe cozinheire.

-ois/-uas → -ues

- Eles são dois.[34] → Elus são dues.

-e/-a → -e

- Ele é o governante.[35] → Elu é ê governante.

-ço/-ça → -ce

- O Rui é bom moço.[36] → Rui é boe moce.

[30] English: “João is my brother.”
[31] English: “Lúcia is a bully.”
[32] English: “Mariana is an actress”
[33] English: “Dani is a good cook.”
[34] English: “There are two of them.”
[35] English: “He is the housekeeper.”
[36] English: “Rui is a good guy.”

B.2.4.A Termos que já são género-neutros / Terms that are already gender-neutral

PT Formas que já são género-neutras mantém-se.

EN Terms which are already gender-neutral are not rewritten.

- O Baltasar é brilhante.[37] → Baltasar é brilhante.
- A Mariana é socialista.[38] → Mariana é socialista.

PT É de notar que nem todos os termos existentes género-neutros necessitam de concordância.

EN It might be worth to note that not all existing gender-neutral terms require gender agreement.

- O Aylton é um indivíduo único.[39] → Aylton é um indivíduo único.
- A Mónica é uma criança.[40] → Mónica é uma criança.

PT Quando os termos necessitam de concordância, esta é feita de acordo com as regras para a respetiva classe de palavras.

EN When terms require gender agreement, this is done according to the rules for the respective word class.

- A Joana é uma boa presidente.[41] → Joana é ume boe presidente.
- O Pedro é o melhor aluno.[42] → Pedro é ê melhor estudante.

B.2.4.B Uso de Sinónimos / Synonym Usage

PT Em instâncias onde existem sinónimos género-neutros de um termo, estes devem ser preferidos à utilização das regras anteriores.

[37] English: “Baltasar is brilliant.”
[38] English: “Mariana is a socialist.”
[39] English: “Aylton is an unique individual.”
[40] English: “Mónica is a child.”
[41] English: “Joana is a good president”
[42] English: “Pedro is the best student.”

EN In instances where there are gender-neutral synonyms for a term, these should be used instead of the above rules.

- A Luísa é professora.[43] → Luísa é docente.
- O Fred é marido do Carlos.[44] → Fred é cônjuge de Carlos.
- A Paula é uma excelente mulher.[45] → Paula é uma excelente pessoa.

[43] English: “Luísa is a teacher.”
[44] English: “Fred is Carlos’s husband.”
[45] English: “Paula is an excelent woman.”

C

LLM Prompting

Prompt "Generate 5 examples of Portuguese gender-neutral sentences, following the "elu" system. Bear in mind that, in the "elu" system, terms that refer to humans tend to have an -e termination. Make sure to use sentences that include nouns, adjectives, and verbs."

Generated Sentences Wrong gender-neutral terms or terms whose gender agreement is not correct are marked in bold.

- Elu é **um** estudante dedicade. (They are a dedicated student.)
- Elu adora explorar novos lugares e culturas. (They love exploring new places and cultures.)
- Elu trabalha como **profissionale** da área de tecnologia. (They work as a technology professional.)
- Elu é uma pessoa **criative** e **talentose**. (They are a creative and talented person.)
- Elu está **escrevende** um livro incrível. (They are writing an amazing book.)