

Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender

Anne Lauscher
MilaNLP
Università Luigi Bocconi
Milan, Italy
anne.lauscher@unibocconi.it

Archie Crowley
Linguistics
University of South Carolina
Columbia, SC, USA
acrowley@sc.edu

Dirk Hovy
MilaNLP
Università Luigi Bocconi
Milan, Italy
dirk.hovy@unibocconi.it

Abstract

Trigger warning: This paper contains examples that contain transphobic language.

The world of pronouns is changing – from a closed word class with few members to an open set of terms to reflect identities. However, Natural Language Processing (NLP) barely reflects this linguistic shift, resulting in the possible exclusion of non-binary users, even though recent work outlined the harms of gender-exclusive language technology. The current modeling of 3rd person pronouns is particularly problematic. It largely ignores various phenomena like neopronouns, i.e., novel pronoun sets that are not (yet) widely established. This omission contributes to the discrimination of marginalized and underrepresented groups, e.g., non-binary individuals. It thus prevents gender equality, one of the UN’s sustainable development goals (goal 5). Further, other identity-expressions beyond gender are ignored by current NLP technology. This paper provides an overview of 3rd person pronoun issues for NLP. Based on our observations and ethical considerations, we define a series of five desiderata for modeling pronouns in language technology, which we validate through a survey. We evaluate existing and novel modeling approaches w.r.t. these desiderata qualitatively and quantify the impact of a more discrimination-free approach on an established benchmark dataset.

1 Introduction

Pronouns are an essential component of many languages and often one of the most frequent word classes. Accordingly, NLP has long studied tasks related to them, e.g., pronoun resolution (e.g., Hobbs, 1978). Simplistically, they can be defined as “a word (such as I, he, she, you, it, we, or they) that is used instead of a noun or noun phrase”.¹

¹Essential definition provided by the Merriam Webster Online Dictionary at <https://www.merriam-webster.com/dictionary/pronoun>











Nom.	Acc.	Poss. (dep.)	Poss. (indep.)	Reflexive
<i>Gendered Pronouns</i>				
he	him	his	his	himself
she	her	her	hers	herself
<i>Gender-Neutral Pronouns</i>				
they	them	their	theirs	themselves
<i>Neopronouns</i>				
thon	thon	thons	thons	thonselves
e	em	es	ems	emself
xe	xem	xyr	xyr	xemself
ey	em	eir	eirs	emself
e	em	eir	eirs	emself
ze	zir	zir	zirs	zirself
...				
<i>Nounself Pronouns</i>				
star	star	stars	stars	starself
vam	vamp	vamps	vamps	vampself
...				
<i>Emojiself Pronouns</i>				
		 s	 s	 self
		 s	 s	 self
...				
<i>Numberself Pronouns</i>				
0	0	0s	0s	0self
$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$ s	$\frac{1}{3}$ s	$\frac{1}{3}$ self
...				
<i>Nameself Pronouns</i>				
John	John	Johns	Johns	Johnself
...				

Table 1: Non-exhaustive overview of phenomena related to third-person pronoun usage in English.

Linguistic studies have pointed out the complexity of pronouns, though (e.g., Postal et al., 1969; McKay, 1993). Pronouns can carry demographic information – in English, for example, information about the *number* of referents and a single referent’s (grammatical) *gender*.² Pronouns can convey even more information in other, non-pro-drop languages.³ Consider Arabana-Wangkangurru, a

²Grammatical and social gender should not be confounded, but are often treated interchangeably by lay audiences.

³First described by Perlmutter (1968), the phenomenon of “pro-drop languages” relates to languages, in which, in certain cases, some classes of pronouns can be omitted (e.g., Italian).

language spoken in Australia, in which a speaker uses different pronouns depending on whether the referent is part of the same social or ritual group (moiety) (Hercus, 1994). As such, pronouns shape how we perceive individuals and can even reflect cultural aspects (e.g., Kashima and Kashima, 1998) and ideologies (e.g., Muqit, 2012). Consequently, pronoun usage should be considered a sensitive aspect of natural language use.

Accordingly, in many western societies, these phenomena have been drawing more and more attention. For instance, in 2020, the American Dialect Society voted “(My) Pronouns” as the 2019 Word of the Year and *Singular “They”* as the Word of the Decade (Roberts, 2020). Recently, there has been a shift in pronoun usage (Krauthamer, 2021), partially due to shifts in the perception of *gender*, driven by the queer-feminist discourse (e.g., Butler, 1990, 2004). Related to this is the open discussion of identity beyond binary gender. For instance, a person who does not identify their gender within the gender binary (e.g., a *nonbinary* or *genderqueer* person) might use *singular “they”* as their pronoun. Recently, the French dictionary “Le Robert” added the non-binary pronoun “*iel*” to its list of words.⁴

This “social push” to respect diverse gender identities also affects NLP. Recent studies have pointed out the harms of the current lack of non-binary representation in data, embeddings, and tasks (Cao and Daumé III, 2021; Dev et al., 2021). However, the research landscape on modern pronoun usage is surprisingly scarce, hindering progress for fair and inclusive NLP. This lacuna is in direct contradiction of the UN’s sustainable development goals,⁵ which include gender equality (goal 5).

Linguistic research has identified further identity aspects of pronouns, beyond gender (Miltersen, 2016). Specifically, *nounself* pronouns are functionally turning pronouns from a *closed* into an *open* word class. To the best of our knowledge, NLP has completely ignored these aspects so far. We did not find a single work systematically describing all of the currently existing phenomena, even just in English 3rd person pronoun usage (let

alone other languages).⁶ In contrast, many discussions are taking place on queer Wikis and forums. While it is still unclear which of these phenomena will persist over the following decades, people use and discuss them. Accordingly, we as a research community should adapt.

Contributions. **1)** We are the first to provide a systematic overview of existing phenomena in English 3rd person pronoun usage. Our results will inform future NLP research on ethical NLP and non-binary representation. We provide the first NLP work acknowledging *otherkin* identities. We support our observations with a corpus analysis on Reddit. **2)** Based on our overview, we derive five desiderata for modeling third-person pronouns, which we validate with a survey among 39 individuals (coupled with a pre-study with 149 participants), most of which identify as non-binary. Based on these criteria, **3)** we discuss various existing and novel paradigms for *when* and *how* to model pronouns in NLP. **4)** Finally, we quantify the impact of discrimination-free non-modeling of pronouns on a widely established benchmark.

2 Related Work

While there are some works in NLP on gender-inclusion (e.g., Dev et al., 2021) and gender bias in static (e.g., Bolukbasi et al., 2016; Gonen and Goldberg, 2019; Lauscher and Glavaš, 2019; Lauscher et al., 2020, *inter alia*) and contextualized (e.g., Kurita et al., 2019; Bordia and Bowman, 2019; Lauscher et al., 2021, *inter alia*) language representations as well as works focusing on specific gender bias in downstream tasks (e.g., Rudinger et al., 2018; Webster et al., 2018; Dev et al., 2020; Barik-eri et al., 2021), we are not aware of any work that deals with the broader field of *identity-inclusion*. Thus, there is no other NLP work that deals with a larger variety of pronouns and acknowledges pronouns as an open word class. For surveys on the general topic of unfair bias in NLP we refer to Blodgett et al. (2020) and Shah et al. (2020). Recently, Dev et al. (2021) pointed at the representational and allocational harms (Barocas et al., 2017) arising from gender-exclusivity in NLP. They surveyed queer individuals and assessed non-binary representations in existing data sets and embeddings. In contrast, we specifically look at third-person pronoun usage and how to model such phenomena. Webster

In contrast, in “non-pro-drop languages”, pronouns cannot be omitted (e.g., German).

⁴<https://dictionnaire.lerobert.com/di-s-moi-robert/raconte-moi-robert/mot-jour/pourquoi-le-robert-a-t-il-integre-le-mot-iel-dans-son-dictionnaire-en-ligne.html>

⁵<https://sdgs.un.org/goals>

⁶For instance, while we found hits for the Google Scholar query “*neopronoun*”, we did not get any results for variants of “*nameself pronoun*”, or “*emojiself pronoun*”.

et al. (2018) provide a balanced co-reference resolution corpus with a focus on the fair distribution of pronouns but only focus on the gendered binary case. Closest to us, Cao and Daumé III (2021) discuss gender inclusion throughout the NLP pipeline beyond binary gender. While they are the first to consider gender-neutral pronouns, including some neopronouns, they do not acknowledge the broader spectrum of identity-related pronoun phenomena.

3 A Note on Identity and Pronouns

This work focuses on the relationship between identity and pronouns. *Identity* refers to an individual’s self-conceptualization, relating to the question of what makes each of us unique (Maalouf, 2000). It can be seen as a two-way process between an individual and others (Grandstrand, 1998), and relates to different dimensions, e.g., one’s gender.

Gender Identity. Gender identity, as opposed to gender expression or sex, is one’s inner sense of gender (Stryker, 2017; Keyes et al., 2021). In this work, we recognize gender identities beyond a cisnormative binary (*cis man*, *cis woman*), e.g., transgender, non-binary, agender, etc.

Otherkin Identity. Individuals with otherkin identity do not entirely identify as human (Laycock, 2012), e.g., vamp. Miltersen (2016) note that otherkin individuals often identify with *nounself* pronouns matching their kin.

Stryker (2017) highlights the strong relationship between gender identity and pronouns. As Raymond (2016) notes, pronoun choices construct the individual’s identity in conversations and the relationship between interlocutors. According to Cao and Daumé III (2021), pronouns are a way of expressing referential gender. Referring to an individual with sets of pronouns they do not identify with, e.g., resulting in misgendering, is considered harmful (e.g., Dev et al., 2021).

4 Phenomena in Third-person Pronoun-Usage

We describe existing phenomena and analyze their presence in a collection of threads from Reddit.⁷

4.1 Existing Phenomena

Overall, individuals can choose n sets of pronouns with $n \geq 0$. If $n = 0$, the individual does not identify with any singular 3rd person pronoun. If $n > 1$, the individual identifies with more than

⁷<https://www.reddit.com>

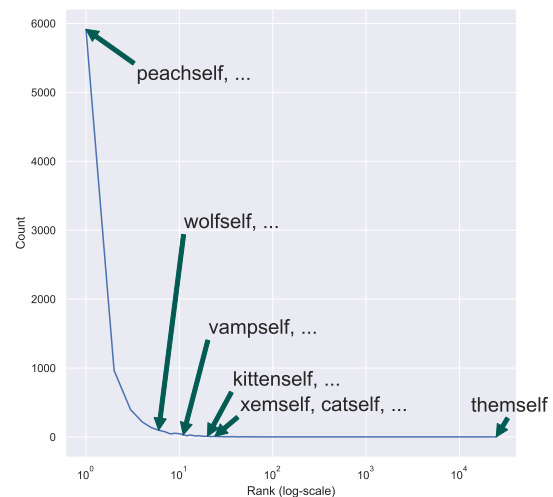


Figure 1: Token ranks (log-scale) and rank counts of the tokens returned against our reflexive regular expression pattern from Reddit with example annotations.

one set of pronouns. Each set is possibly reflecting overlapping or non-overlapping aspects of their identity. We provide examples of these sets in Table 1. Note that this list is non-exhaustive and that the described phenomena are non-exclusive.

Gendered Pronouns. In English, two sets of gendered pronouns are available, *he/him/himself* and *she/her/herself*.

Gender-Neutral Pronouns. Given the history of generic singular *they* in English (e.g., Who was at the door? *They* left a note.), there has been an uptake of singular *they* by non-binary individuals as a gender-neutral pronoun option⁸ (Conrod, 2019; Konnelly and Cowper, 2020). Further, there has been increasing institutional recognition with dictionaries and style guides supporting its use.

Neopronouns. As an alternative to the singular *they*, individuals started creating and sharing novel sets of 3rd person pronouns (McGaughey, 2020). More traditional and rather well-known sets of neopronouns include, e.g., the so-called Spivak pronouns *e/emself* (used in (Spivak, 1990)) and related variations. During our research, we observed various subcategories of neopronouns, only partially described in the academic literature.

Nounself Pronouns. Nounself pronouns are “[...] prototypically transparently derived from a spe-

⁸<https://gendercensus.com/results/2021-worldwide-summary/>

cific word, usually a noun” (Miltersen, 2016). Individuals may identify with certain nouns, possibly corresponding to aspects of their identity, e.g., *kitten/kittenself*, *vamp/vampself*. The author notes the difficulty of clearly defining nounself pronouns, neopronouns, and other phenomena. The phenomenon is assumed to have first appeared in 2013.

Emojiself Pronouns. Similar to nounself pronouns, individuals may identify with sets of emojis, possibly reflecting different aspects of their identity, e.g., 🥰/🥰self. Emojiself pronouns are intended for written communication. Note that, at the time of writing this manuscript, no academic description of emojisself pronouns seems to exist. However, we were able to find evidence of their existence on several social media platforms and wikis, e.g., Tumblr,⁹ MOGAI Wiki,¹⁰ Twitter,¹¹ and Reddit.¹²

Numberself Pronouns. Another form of neopronouns/ nounself pronouns are numberself pronouns. Analogous to before, we assume that here, the individual identifies or partially identified with a number, e.g., *0/0/0s/0s/0self*.¹³

Nameself Pronouns. Individuals may identify with pronouns build from their name, e.g., *John/Johnself*, overlapping with nullpronomials.¹⁴

Alternating Pronouns. Suppose someone identifies with more than one set of pronouns. In that case, the pronouns they identify with can be either equally identified-with sets, or potentially change depending on the context (*mutopronominal*). As such, individuals who are also performers may use *stage pronouns*. Similarly, genderfluid individuals may identify with a certain pronoun at a certain point in time (*pronoun fluidity*, (Cherry-Reid, 2020)). Some individuals identify with the pronouns of the person who is referring to them (*mirrored pronouns*). Others use set(s) of *auxiliary*

pronouns, e.g., for situations when people referring to them have problems with using the most identified-with sets of pronouns (e.g., in the case of emojisself pronouns and oral communication). Alternating pronoun sets may be even used in the same sentence for the same individual.¹⁵

No Pronouns. Some individuals do not identify with any pronouns. In this case, some individuals identify most with their name being used to refer to them, nameself pronouns, or avoid pronouns.

4.2 Corpus Analysis: Neopronouns in Reddit

Setup. We conduct an additional analysis for the presence of neopronouns in Reddit. To this end, we use Reddit threads (2010–2021), cleaned by previous work and provided through Huggingface Datasets (127,445,911 lines). The data set includes comment title, text, and subreddit.¹⁶ Since we are interested in capturing novel pronouns, but the list of possible pronouns is indefinite, we proxy neopronouns via the suffixes *self* and *selves* to indicate the reflexive case. We match them through a regular expression. Additionally, we filter out non-3rd person pronouns (e.g., *yourself*, *ourselves*, plural *themselves*) as well as common variations of these (e.g., *urself*) and other common non-pronoun expressions we found in the data (e.g., *do-it-yourself*). This process leaves us with a total of 9,075 unique tokens with in total 74,768 textual mentions.

Results. Unsurprisingly, an initial manual analysis reveals that many of the matches are false positives, i.e., not real neopronouns like *non-self*, a common concept in Buddhist philosophy. However, our method still finds relevant cases. Note, that in this work, we do not explicitly deal with false positives – we are merely interested in whether our heuristic helps us to detect some sets of neopronouns at all, thus demonstrating their existence in real-world conversations. Examples of neopronouns we found are depicted in Table 2. Many discussions containing nounself pronouns center on the phenomena themselves, including, e.g., individuals stating that they are interested in using a specific pronoun or that they refuse to acknowledge the phenomenon. Some discussions involve people reporting on personal experiences and problems and seeking advice. To obtain a high-level quantitative

⁹E.g., <https://pronoun-archive.tumblr.com/post/188520170831>

¹⁰<https://mogai.miraheze.org/wiki/Emojiself>; according to the article, the origin of emojisself pronouns is unclear but might date back to 2017

¹¹Example of a user complaining about LinkedIn not allowing for emojisself pronouns in the pronoun field: <https://twitter.com/frozenpandaman/status/1412314202119700480/photo/1>

¹²E.g., https://www.reddit.com/r/QueerVexilology/comments/p09nek/i_made_a_flag_for_the_emojiself_pronoun_set/

¹³<https://pronoun-provider.tumblr.com/post/148452374817/i-think-numbers-as-pronouns-would-be-pretty-cool>

¹⁴https://pronoun.fandom.com/wiki/Null_pronominal

¹⁵https://www.reddit.com/r/NonBinary/comments/jasv5r/alternating_pronouns_in_sentence/

¹⁶<https://huggingface.co/datasets/sentence-transformers/reddit-title-body>

Match	Thread Title	Thread Excerpt
meowself	<i>Fureedom Mewnite can die in my litterbox.</i> <i>Neopronouns are going too far.</i>	<i>I don't like this game. But I still want meowself to play it, meow. Cause it's fun, even though I hate it.</i> <i>I get some pronouns like ze/zir, xe/xem, etc. I agree with those. But why are people using ghost/ghostself and meow/meowself? That's really utter bullshit.</i>
bunself	<i>I am genderfluid, and pansexual. I have a lot of SJW friends. AMA!</i> <i>Xi am so proud to announce that the new word of the year is.....</i>	<i>They/them pronouns are coolest with me, but I won't be angry if you use he or she. You can use bun/buns/bunself, if you are feeling special. (That's a joke.)</i> <i>–Cinnagender– which means you identify with our beloved and innocent cinnamon buns. The pronoun set is cinne/cinns/cinnself or alternatively bun/buns/bunself i am so happy to be a member of a community that ignores the oppressive gender binary, which is a social construct, i.e., it is not real</i>
zirself	<i>Ran into our first roadblock</i> <i>If you're a horrible person online, you're probably a horrible person offline too.</i>	<i>I asked what I could do to help zir lowering the feeling of dysphoria, and ze said zed maybe feel better about zirself if zed drink a tea.</i> <i>Hello folks. Omg. I think this individual is about to hurt zirself! (emphasis on "zirself". COMEDIC GENIUS)</i>

Table 2: Example neopronouns and corresponding excerpts from Reddit retrieved via our heuristic method. We slightly modified the excerpts to lower searchability and increase the privacy of the users.

view, we compute the matches’ ranks as the number of texts in which particular matches occurred (including false positives) against their number of tokens (e.g., there is only 1 match, which appeared in 24,697 texts; there are 2 matches, which appeared in 198 texts, etc.). We show the results in Figure 1 (log-scale). The result is a highly skewed Zipf’s distribution: while the highest ranks appear only once (e.g., *themselves* with 24,697 mentions), some tokens appear only a couple of times (e.g., the neopronoun *xemself* with 24 mentions), and the vast majority appears only once (e.g., many nounself pronouns such as *peachself*).

5 How Can and Should We Model Pronouns?

We devise five desiderata based on our observations, personal experiences, expert knowledge from interactions with LGBTQIA+ associates, and informal discussions with individuals using gender-neutral pronouns. We validate the desiderata through a survey. Here, we collect opinions from 39 individuals (149 in the pre-study), most of whom identify as non-binary. We then assess how well classic and novel NLP pronoun modeling paradigms fulfill the criteria.

5.1 Desiderata

D1. Refrain from assuming an individual’s identity and pronouns. A model should not assume an individual’s identity, e.g., gender, or pronouns based on, e.g., statistical cues about an individ-

ual’s name, also not in a binary gender setup. Only because the name *John* typically appears together with the pronoun *he*, the model should not assume that a person with the name *John* identifies as a man and that every *John* uses the pronoun *he*.

D2. Allow for the existing sets of pronouns as well as for neopronouns. A model should be able to handle not only the existing set of “standard” pronouns in a language but also other existing pronouns, e.g., neopronouns.

D3. Allow for novel pronouns at any point in time. On top of D2, a model should allow for novel, i.e., unseen, pronouns to appear at any point in time. This condition is necessary to account for the fact that neopronouns are not a fixed set, but evolving, and because related phenomena (emojiself and nameself pronouns) turn pronouns from a *closed* to an *open class* part of speech.

D4. Allow for multiple, alternating, and changing pronouns. A model should not assume that the pronoun set for an individual at time t will be the same as at time $t - 1$. Even within the same sequence, pronoun sets might change.

D5. Provide an option for individuals to define their sets of pronouns. While most NLP models are trained offline and do not interact with the user, some are designed to interact with individuals, e.g., dialog systems. Here, letting individuals provide their sets of pronouns can help avoid harmful interactions (depending on the sociotechnical scenario).

5.2 Validation

Survey Design. We divide the survey into three parts: first, participants are asked for demographic information (age, (gender) identity, native language(s), pronouns). The second part asks for their opinion on D1–D5. We first provide a general contextualization of our research and describe the task. Participants are asked to indicate how much they agree with each desideratum (ordinal scale, 1 (*I don’t agree*) to 5 (*I absolutely agree*)). We also allow for leaving additional comments. The third part relates to a case study on machine translation. We inform the participants that their participation is completely voluntary and that they will not receive any compensation. All questions are optional. To avoid sequence effects, we create multiple versions of the survey shuffling the order of the desiderata. We obtained ethical approval for the design by one of our universities’ institutional review board.

Survey Distribution. Opting to collect opinions from affected individuals, we distribute the survey through various international LGBTQIA+ networks, e.g., QueerInAI,¹⁷ Committee on LGBTQ+[Z] Issues in Linguistics,¹⁸ as well as through local LGBTQIA+ groups, e.g., Transgender Network Switzerland.¹⁹ In an initial pre-study, which was open for participation between 22nd of March and 4th of May 2022, we validated our design. In total, 149 individuals participated in this phase (more than 8x more than in (Dev et al., 2021)).²⁰ The main phase of the survey was open for participation between 18th of June and 1st of August 2022.

Participant Statistics. In total, 44 individuals participated in the main phase of our survey, more than in any other survey on (gender) identity and language technology we are aware of. Participant ages range from 14 to 43 (the majority between 20 and 30). For the rest of the analysis, we removed all records from individuals under the age of 18. These individuals indicated that they speak diverse native languages (e.g., German, English, Danish, Persian, Russian). Participants provided between 0 and 4 identity terms (e.g., genderfluid, genderqueer, trans*masculine, etc.), with the majority identify-

¹⁷<https://sites.google.com/view/queer-in-ai>

¹⁸<https://www.linguisticsociety.org/content/committee-lgbtq-z-issues-linguistics-cozil>

¹⁹<https://www.tgns.ch>

²⁰All trends observed in the pre-study were confirmed in the main phase of the survey.

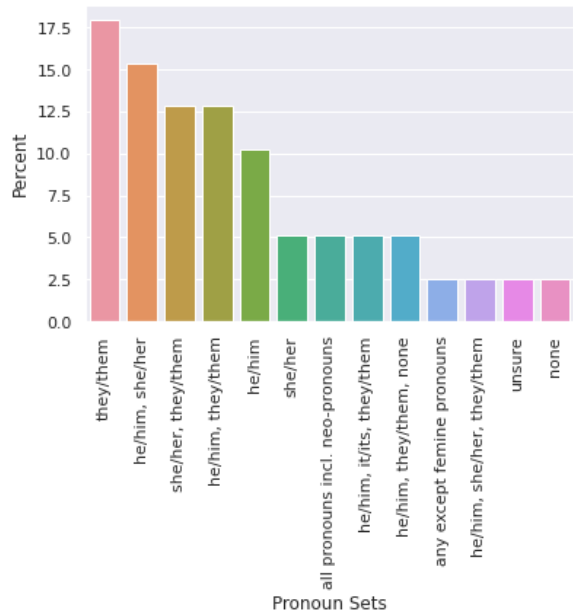


Figure 2: Distribution over English pronoun sets the participants of our survey identify with. We apply slight lexical normalization to the pronoun strings provided in the free text field (e.g., *he* → *he/him*, etc.).

Score	1	2	3	4	5	Avg
D1	5.13%	2.56%	0.00%	33.33%	58.97%	4.38
D2	2.56%	0.00%	5.13%	17.95%	74.36%	4.62
D3	2.56%	0.00%	12.82%	25.64%	58.97%	4.38
D4	2.56%	5.13%	2.56%	23.08%	66.67%	4.46
D5	0.00%	2.56%	5.13%	20.51%	71.79%	4.62

Table 3: Agreement distribution for our desiderata D1–D5. We report the % agreement per score (1=*I don’t agree*, 5=*I absolutely agree*) as well as the average score.

ing as non-binary. Thus, we believe our results to reflect cultural and (gender) identity diversity. Figure 2 shows the distribution over the English pronoun sets participants identify with.

Agreement with the Desiderata. We show the distribution of agreement scores for each desideratum in Table 3. Overall, we note high agreement (on average 4.38–4.62). Thus, we conclude our proposed desiderata to provide valuable orientation for the treatment of pronouns leading to more identity-inclusive language technology.

5.3 Modeling Paradigms

We compare four general modeling paradigms with the desiderata D1–D5 in Table 4.

Classic Statistical Modeling. Traditionally, pronouns have been treated as a *closed word class*. Generally, statistical models do not make as-

Paradigm	D1	D2	D3	D4	D5
Classic	✗	✗	✗	✗	✗
Bucketing	✗	✓	✓	?	✗
Delexicalization	✓	✓	✓	✓	✗
Post-hoc	✓	✓	✓	✓	✓

Table 4: Modeling paradigms and how they allow for fulfilling the desiderata D1–D5.

	Train	Dev	Test	Total
PRP	64,476	7,881	8,067	80,424
PRP\$	14,535	1,783	1,935	18,253
Total	79,011	9,664	10,002	98,677

Table 5: Number of pronoun replacements in the training, development, and test portion of OntoNotes 5.0 for PRP and PRP\$, respectively.

sumptions about this (except if the vocabulary is manually curated). However, in models exploiting co-occurrences, e.g., via word embeddings (GloVe (Pennington et al., 2014)) or deep language models (BERT (Devlin et al., 2019)), the models will likely misrepresent underrepresented pronoun-related phenomena. Dev et al. (2021) provided an initial insight by showing that singular *they* and the neopronouns *xe* and *ze* do not have meaningful vectors in GloVe and BERT.

Bucketing. One option, previously discussed by Dev et al. (2021), is to apply bucketing, i.e., to decide on a fixed number of majority classes, e.g., *male* pronouns, *female* pronouns, and one or multiple classes for the “rest of the pronouns”, e.g., *other*. The advantage of this approach is that it can map existing and novel pronouns to the *other* class. However, it still makes identity assumptions – and due to unequal representations of *main* and *other* classes, it will inevitably lead to discrimination.

No Modeling – Delexicalization. Given that the classic approach and bucketing both lead to unfair treatment of underrepresented groups, the alternative is to explicitly not model pronouns in their surface forms. This process, commonly named delexicalization, has proved helpful for other tasks where models capture misleading lexical information, e.g., fact verification (e.g., Suntwal et al., 2019), or resource-lean scenarios, e.g., cross-lingual parsing (e.g., McDonald et al., 2011).²¹ In this case, the model is forced to not rely on spurious lexical cues related to gender, e.g., that *John* occurs most often

²¹In fact, accounting for novel pronouns and novel ways of using pronouns is a resource-lean scenario.

with the pronoun *he*. Instead, the model learns a single representation for all pronouns and relies on other task-related conceptual and commonsense information for disambiguation.

Post-hoc Injection of Modeling Information/Modeling at Test Time.

For human-to-human interactions, several LGBTQIA+ guides recommend to (1) first try generic pronouns (e.g., singular *they*), and (2) switch to other sets of pronouns once the conversation partner communicates them. For uncommon or novel pronouns, several web pages have explicitly been set up for practicing how to use them.²² In this work, *we propose that NLP systems should work similarly* – if technically possible and depending on the concrete sociotechnical deployment scenario. To this end, we can use intermediate training procedures (e.g., Hung et al., 2021) for pronoun-related model refinement. E.g., we can use synthetic data created through similar procedures as the ones employed on these websites. Another option is only model pronouns at test time, e.g., through simple replacement procedures.

6 The Effect of Delexicalization

In §5.3, we discussed delexicalization, i.e., not modeling lexical forms of pronouns, as one way to counter exclusion in statistical modeling and bucketing. A possible counter-argument against this approach is that omitting the surface forms will lead to poor performance on pronoun-related tasks. We experimentally quantify the loss from (fairer) delexicalization in co-reference resolution.

6.1 Experimental Setup

Task, Dataset, and Measures. We resort to co-reference resolution, a task where knowledge about pronouns and related gender assumptions play an essential role. We use the English portion of *OntoNotes 5.0* (Weischedel et al., 2012), which consists of texts annotated across five domains (news, conversational telephone speech, weblogs, USENET newsgroups, broadcast, and talk shows). We prepare three variants: (i) the original data; (ii) we *replace all pronouns* in the test set with the respective part-of-speech token, according to the Penn Treebank Project (Santorini, 1990), i.e., PRP for personal pronouns, and PRP\$ for possessive pronouns. Finally, we provide a version (iii) where we replace pronouns in all splits. Note that our

²²E.g., https://www.practicewithpronouns.com/#/?_k=66emp7

	MUC			CEAF $_{\phi_4}$			B ³			AVG		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
(Dobrovolskii, 2021)	84.9	87.9	86.3	76.1	77.1	76.6	77.4	82.6	79.9	–	–	81.0
- reproduction	84.7	87.5	86.1	75.6	76.7	76.1	77.2	82.0	79.5	79.2	82.1	80.6
- replace test set	69.7	70.7	70.2	63.2	49.1	55.2	50.1	56.1	52.9	61.0	58.6	59.4
$\Delta_{repl.test-repr.}$	-15.0	-16.8	-15.9	-12.4	-27.6	-20.9	-27.1	-25.9	-26.6	-18.2	-23.5	-21.2
- replace all	81.6	83.1	82.4	73.08	72.9	73.0	72.3	75.3	73.7	75.7	77.1	76.4
$\Delta_{repl.all-repr.}$	-3.1	-4.4	-3.7	-2.5	-3.8	-3.1	-4.9	-6.7	-5.8	-3.5	-5.0	-4.2

Table 6: Results of the delexicalization experiment. We report the results of the RoBERTa large-based word-level co-reference resolution model from Dobrovolskii (2021), our reproduction, and variants trained or tested on versions of the data set in which we replace the pronouns. All scores were produced using the official CoNLL-2012 scorer. We report precision (P), recall (R), and F1-score (F1) for MUC, CEAF $_{\phi_4}$, and B³, respectively, as well as the averages (AVG). The rows highlighted in gray indicate the obtained losses.

strategy is pessimistic as we also replace non-3rd person pronouns, i.e., *I*, *you*, etc. We show the number of replacements in Table 5. For scoring, we use the official CoNLL-2012 scorer (Pradhan et al., 2012). We report the results in terms of MUC (Vilain et al., 1995), B³ (Bagga and Baldwin, 1998), and CEAF $_{\phi_4}$ (Luo, 2005) precision, recall, and F1-measure, and the averages across these scores.

Models and Baselines. We want to obtain an intuition about the tradeoffs in the delexicalization setup, not to outperform previous results. For this reason, we resort to the recently proposed word-level co-reference model (Dobrovolskii, 2021), a highly efficient model competitive with the state-of-the-art. The model consists of a separate co-reference resolution module and a separate span extraction module. In an initial step, we compute token representations from a Transformer (Vaswani et al., 2017)-based encoder through the aggregation of initial representations via learnable weights. Next, we compute co-reference relationships. We pass the token representations into an antecedent-pruning procedure based on a bilinear scoring function. This way, we obtain k antecedent candidates for each token through coarse-grained scoring. An additional feed-forward neural network computes finer-grained scores. The final antecedent score is the sum of those two scores. Finally, we select the candidate with the highest score as the antecedent. Negative scores indicate no antecedent. Tokens assumed to be part of a co-reference relationship are passed into the span extraction module. The module consists of an additional feed-forward network, followed by convolutions with two output channels (for start and end scores). For further details, see the original work. Our baseline is the model trained and evaluated on the original OntoNotes portions

(*reproduction*). We compare with the evaluation of this model on the pronoun-replaced test set (*replace test set*) and a version of this model trained on the replaced training set and evaluated on the replaced test set, respectively (*replace all*).

Model Configuration, Training, and Optimization. We choose RoBERTa large (Liu et al., 2019)²³ as the base encoder and fix all other hyperparameters to the ones provided in the original implementation of Dobrovolskii (2021): the window size is set to 512 tokens, dropout rate to 0.3, the learning rate of the encoder is set to $1 \cdot 10^{-5}$ and of the task-specific layers to $3 \cdot 10^{-4}$, respectively. We train the co-reference module with a combination of the negative log marginal likelihood and binary cross-entropy as an additional regularization factor (weight set to 0.5). The span extraction module is trained using cross-entropy loss. We optimize the sum of the two losses jointly with Adam (Kingma and Ba, 2015) for 20 epochs and apply early stopping based on validation set performance (word-level F1) with a patience of 3 epochs.

6.2 Results and Discussion

We show the results in Table 6. We are roughly able to reproduce the results reported by (Dobrovolskii, 2021), confirming the effectiveness of their approach and the validity of our setup. When we replace pronouns in the test set, the results drop massively, with up to -27.6 percentage points CEAF $_{\phi_4}$ recall. This decrease demonstrates the heavy reliance of this model on the lexical surface forms of the pronoun sets seen in the training. However, when we replace the pronouns in the training portion of OntoNotes with the special tokens, we can mitigate these losses by a large margin (losses up

²³<https://huggingface.co/roberta-large>

to $-5.8 B^3$ F1, and on average -4.2 F1). These results are highly encouraging, given that a) we replaced *all* pronouns, including non-third person pronouns, and b) the model has not been trained on these placeholders in the pretraining phase. The model can not rely on possibly discriminating correlations between names or occupations and pronoun sets. It therefore represents neopronouns the same way as established pronoun sets. A delexicalization approach can increase fairness in co-reference resolution and retain high system performance. We can expect even smaller drops from a more careful selection of replacements, and, possibly, from intermediate training procedures that strengthen the representation of the placeholder tokens.

7 Conclusion

This work provides an initial overview of the plethora of current phenomena in 3rd person pronoun usage in the English language. For practical and ethical reasons, the NLP community should acknowledge the broad spectrum of possible identities and the respective manifestations in written and oral communication. Language is consistently evolving, and NLP researchers and practitioners should account for this to provide genuinely inclusive systems. Notably, pronouns, traditionally handled as a closed class of words, seem to function closer to an open class. Based on the observations from our literature search, research in non-academic, publicly-available writing, a corpus study, and a survey, we defined five desiderata. We validated those and applied them to the discussion of existing and novel modeling paradigms. Our findings raise the questions *when* and *how* to model pronouns and whether and how to *include users* in these decisions. With this work, we hope to start a broader discussion on the topic. Our study can inform future NLP research and serve as a starting point for creating novel modeling procedures. All code needed to reproduce our experiments is publicly available at <https://github.com/anlausch/pronouns>.

Acknowledgments

The work of Anne Lauscher and Dirk Hovy is funded by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR). We thank Emily Bender, Chandler May, and Arjun Subramonian for sharing their ideas related to our project.

Further Ethical Discussion

We have described phenomena related to third-person pronouns focusing exclusively on the English language. Naturally, this work comes with several limitations. For instance, while we pointed the reader to the variety of pronoun-related phenomena in other languages, a thorough *multilingual and cross-lingual discussion* would have exceeded the scope of this manuscript. This lacuna includes the discussion of neopronouns in other languages. Similarly, while we acknowledge identities beyond binary gender and otherkin identities, due to our focus on pronouns, we did not investigate *other identity-related terms*. This aspect includes their handling in language technology and the range of issues related to identity-exclusivity.

References

- Amit Bagga and Breck Baldwin. 1998. [Algorithms for scoring coreference chains](#). In *Proc. Linguistic Coreference Workshop at the first Conf. on Language Resources and Evaluation (LREC)*, pages 563–566, Granada, Spain.
- Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. [RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.
- Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. [Language \(technology\) is power: A critical survey of “bias” in NLP](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
- Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. [Man is to computer programmer as woman is to homemaker? debiasing word embeddings](#). In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.
- Shikha Bordia and Samuel R. Bowman. 2019. [Identifying and reducing gender bias in word-level language](#)

- models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.
- Judith Butler. 1990. *Gender trouble*, 1st edition. Routledge Classics, New York, NY, USA.
- Judith Butler. 2004. *Undoing Gender*, 1st edition. Routledge, New York, NY, USA.
- Yang Trista Cao and Hal Daumé III. 2021. [Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle*](#). *Computational Linguistics*, 47(3):615–661.
- Katharine A Cherry-Reid. 2020. *Music to Our Ears: Using a Queer Folk Song Pedagogy to do Gender and Sexuality Education*. Ph.D. thesis, University of Toronto (Canada).
- Kirby Conrod. 2019. *Pronouns Raising and Emerging*. Ph.D. thesis, University of Washington.
- Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikrumar. 2020. [On measuring and mitigating biased inferences of word embeddings](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7659–7666. AAAI Press.
- Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. [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*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Vladimir Dobrovolskii. 2021. [Word-level coreference resolution](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hila Gonen and Yoav Goldberg. 2019. [Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ove Grandstrand. 1998. [Identity and deception in the virtual community](#). In Peter Kollock and Marc Smith, editors, *Communities in Cyberspace*, 1st edition, chapter 2. Routledge, London, UK.
- Luise A Hercus. 1994. *A grammar of the Arabana-Wangkangurru language, Lake Eyre Basin, South Australia (Pacific linguistics. Series C)*, 1st edition. Dept. of Linguistics, Research School of Pacific and Asian Studies, Australian National University.
- Jerry R Hobbs. 1978. [Resolving pronoun references](#). *Lingua*, 44(4):311–338.
- Chia-Chien Hung, Anne Lauscher, Simone Paolo Ponzetto, and Goran Glavaš. 2021. [DS-TOD: Efficient domain specialization for task oriented dialog](#). *arXiv preprint arXiv:2110.08395*.
- Emiko S. Kashima and Yoshihisa Kashima. 1998. [Culture and language: The case of cultural dimensions and personal pronoun use](#). *Journal of Cross-Cultural Psychology*, 29(3):461–486.
- Os Keyes, Chandler May, and Annabelle Carrell. 2021. [You keep using that word: Ways of thinking about gender in computing research](#). *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Lex Konnolly and Elizabeth Cowper. 2020. [Gender diversity and morphosyntax: An account of singular they](#). *Glossa: a journal of general linguistics*, 5(1).
- Helene Seltzer Krauthamer. 2021. *The Great Pronoun Shift: The Big Impact of Little Parts of Speech*, 1st edition. Routledge.
- Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. [Measuring bias in contextualized word representations](#). In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.
- Anne Lauscher and Goran Glavaš. 2019. [Are we consistently biased? multidimensional analysis of biases in distributional word vectors](#). In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

- Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2020. [A general framework for implicit and explicit debiasing of distributional word vector spaces](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8131–8138.
- Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. [Sustainable modular debiasing of language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Joseph P Laycock. 2012. “We are spirits of another sort”: Ontological rebellion and religious dimensions of the otherkin community. *Nova Religio: The Journal of Alternative and Emergent Religions*, 15(3):65–90.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Xiaoqiang Luo. 2005. [On coreference resolution performance metrics](#). In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.
- Amin Maalouf. 2000. *On Identity*, 1st edition. Vintage. Translated from French by Barbara Bray.
- Ryan McDonald, Slav Petrov, and Keith Hall. 2011. [Multi-source transfer of delexicalized dependency parsers](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Sebastian McGaughey. 2020. [Understanding neopronouns](#). *The Gay & Lesbian Review Worldwide*, 27(2):27–29.
- John C McKay. 1993. [On the term “pronoun” in italian grammars](#). *Italica*, 70(2):168–181.
- Ehm Hjorth Miltersen. 2016. [Nounself pronouns: 3rd person personal pronouns as identity expression](#). *Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift*, 1(1):37–62.
- Abd Muqit. 2012. [Ideology and power relation reflected in the use of pronoun in osama bin laden’s speech text](#). *International Journal of Social Science and Humanity*, 2(6):557.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- David M Perlmutter. 1968. *Deep and surface structure constraints in syntax*. Ph.D. thesis, Massachusetts Institute of Technology.
- Paul Postal, David A Reibel, and Sanford A Schane. 1969. On so-called pronouns in english. *Readings in English transformational grammar*, pages 12–25.
- Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. [Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes](#). In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40.
- Chase Wesley Raymond. 2016. [Linguistic reference in the negotiation of identity and action: Revisiting the t/v distinction](#). *Language*, 92:636–670.
- Julie Roberts. 2020. 2019 word of the year is “(my) pronouns,” word of the decade is singular “they” as voted by american dialect society. Press Release, American Dialect Society.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. [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)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
- Beatrice Santorini. 1990. [Part-of-speech tagging guidelines for the penn treebank project](#). Technical Report (3rd Version), University of Pennsylvania, School of Engineering and Applied Science.
- Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. [Predictive biases in natural language processing models: A conceptual framework and overview](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.
- Michael Spivak. 1990. *The Joy of TeX: A Gourmet Guide to Typesetting with the AMSTeX Macro Package*, 2nd edition. American Mathematical Society.
- Susan Stryker. 2017. *Transgender history: The roots of today’s revolution*, 2nd edition. Seal Press.
- Sandeep Sunawal, Mithun Paul, Rebecca Sharp, and Mihai Surdeanu. 2019. [On the importance of delexicalization for fact verification](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3413–3418, Hong Kong, China. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural*

Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Marc Vilain, John D Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. [A model-theoretic coreference scoring scheme](#). In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

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

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D Hwang, Claire Bonial, et al. 2012. [Ontonotes release 5.0](#).