A Bayesian account of pronoun and neopronoun acquisition

Cassandra L. Jacobs
Department of Linguistics
State University of New York at Buffalo
Buffalo, NY, USA
cxjacobs@buffalo.edu

Morgan Grobol
MoDyCo
Université Paris Nanterre
Nanterre, France
lgrobol@parisnanterre.fr

Abstract

A major challenge to equity among members of queer communities is the use of one's chosen forms of reference, such as personal names or pronouns. Speakers often dismiss their misuses of pronouns as "unintentional", and claim that their errors reflect many decades of fossilized mainstream language use, as well as attitudes or expectations about the relationship between one's appearance and acceptable forms of reference. We argue for explicitly modeling individual differences in pronoun selection and present a probabilistic graphical modeling approach based on the nested Chinese Restaurant Franchise Process (nCRFP) (Ahmed et al., 2013) to account for flexible pronominal reference such as chosen names and neopronouns while moving beyond form-to-meaning mappings and without lexical co-occurrence statistics to learn referring expressions, as in contemporary language models. We show that such a model can account for variability in how quickly pronouns or names are integrated into symbolic knowledge and can empower computational systems to be both flexible and respectful of queer people with diverse gender expression.

1 Introduction

In contrast to words that are used to label referents as determined by convention (e.g., "cat" refers to CAT-like entities; Brennan and Clark, 1996), people have the autonomy to change their names and update their pronouns to reflect their identity (Zimman, 2019). In many Western cultures, however, personal names and pronouns are usually assigned *to* someone *by* others (e.g., one's parents or the norms of the ambient culture; Lind, 2023), and are highly conventionalized. For example, English canonically has only two animate third-person singular pronouns (i.e., he/him/his and she/her/hers). These pronominal forms as well as personal names are strong cues to gender identity. Within linguistics, this regularity has led to the general practice of treating referring expression generation as a form-to-meaning mapping problem (Enfield and Stivers, 2007). That said, the forms of reference used for someone are neither fixed, nor intrinsic properties of an individual. This paper presents a probabilistic modeling framework that respects a person's right to self-determination (of how to be referred to) without positing form-to-meaning or form-to-feature mappings. Our proposal accounts for the ongoing sociolinguistic change among young Westerners to ask and reinforce their understanding of their peers' self-identities.

The need for modeling pronoun and name use in natural language processing (NLP) is especially important given the increasing prominance of accommodating individuals' identities in the public sphere. Despite major advances in natural language generation, it has proven difficult to incorporate this into modern systems, especially in present-day neural network models. For example, even the most basic rule-based tokenization systems still do not flexibly handle nonbinary forms of address such as "Mx." Furthermore, large language models (LLMs) and commercial generative AI systems perpetuate bias against women and gender minorities by encoding harmful stereotypes in their training data (e.g., negative sentiment; Dev et al., 2021; Ungless et al., 2023) for in marginalized individuals' names, common professions, personal items, and pronouns. This is even more true for queer people outside the gender binary, as datasets regularly exclude nonbinary identities from their construction (Hall et al., 2023; Sakaguchi et al., 2021). Language that does not conform to gender stereotypes is also mishandled by NLP systems (Ghosh and Caliskan, 2023; Havens et al., 2022).

Here, we propose that systems that symbolically encode valid referring expressions for individuals are less prone to these problems. With present limitations in mind, we outline below the basic capabil-

Proceedings of the Queer in AI Workshop, pages 35–40

ities of an ideal system for learning the forms and representations of an individual's referring expressions such as names and pronouns must include:

1. Allow the introduction of new forms into the vocabulary (e.g., novel names or neopronouns)

2. Permit individuals to use a mixture of forms of reference for themselves (e.g., alternating between he/she/they or using different gendered forms in different languages; Moore et al., 2024)

3. Quickly adapt in the face of revision (e.g., updates to a person's name or pronouns), potentially given a single exemplar

4. Allow adaptation to vary by individuals

We further argue that such a system should produce more flexible adaptation for individuals who are more accustomed to such adaptation.

2 A Dirichlet process model of name and pronoun learning

Due to its symbolic nature, our proposed system can learn appropriate forms of address and reference through experience without encoding discriminatory knowledge such as an individual's appearance into their representations. This empowers queer people and supports their autonomy (Lind, 2023; Ovalle et al., 2023; Zimman, 2019). We treat the learning process as the assignment of probabilities of referential forms – pronominal or otherwise – directly to individuals rather than through the medium of individual characteristics (Lauscher et al., 2022).

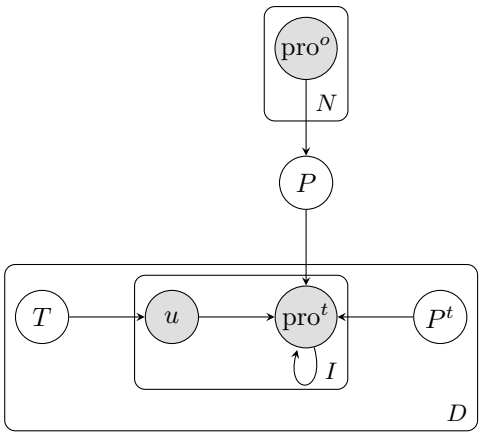

Figure 1: Single speaker model

Latent Dirichlet Allocation (LDA; Blei et al., 2001) is an algorithm that allows the probabilistic assignment of discrete labels (e.g., topics) to collections of events (e.g., documents) on the basis of the contents of the document (e.g., words). As suggested by the name, the topics learned by LDA are latent variables that are unobservable. In this modeling framework, documents are observable objects that are assumed to be generated by sampling words from mixtures of topics. Critically, a trained topic model can be used to estimate what proportion of topics was used to generate that document. These models are in principle infinite, and can have novel topics as well as additional vocabulary items added as a dimension in the vocabulary by trivial extension.

Building on this approach, the nested Chinese Restaurant Franchise Process (nCRFP; Ahmed et al., 2013) allows for models to even learn that different types of documents or users exist. For example, book chapters and magazine articles may have different lexical distributions, and authors within each of those genres may have different lexical preferences. Graphical models have been used to capture variation in language use across different geographical regions (Eisenstein et al., 2010) – analogous to the speaker communities of interest here. Simplified versions of Dirichlet processes (e.g., Beta-binomial priors) have also been applied to learning, as in learning and adaptation to syntactic structures in the context of a conversation (Kleinschmidt et al., 2012).

The present paper expands the metaphor of the nCRFP (Ahmed et al., 2013) to model an individual's learning of referring expressions – and specifically the pronouns – for others. We choose to treat pronouns or similar gender markers as observable objects that have probabilistic assignment to topics (communities of individuals), making pronouns most analogous to words in a document. Furthermore, we can characterize individuals or referents as "documents" that comprise a unique probability distribution over pronouns and names. Extending the metaphor to the hierarchical domain, different communities of learners (topics) may have priors of different strengths and/or more uniform expectations over pronoun use for unfamiliar individuals. Within topics, it is also clear that different groups of learners belong to different communities that reinforce the statistics of use of referring expressions within their communities.

3 Probabilistic graphical model of individual speaker preference

In Figure 1, we present the parametrization of the single-speaker model, which details how a speaker selects pronouns referring to a specific individual t across utterances as a function of their linguistic experience. This model involves the following variables (indices are omitted in the figure for brevity):

$\text{pro}^t_{d,i}$ Produced pronoun referring to t in the interaction i of discourse d. Can be absent, in case where the preferred pronouns are no pronouns. The self-loop allows for both pronoun stability and intentional alternation. That is, speakers can either select a chosen pronoun for a particular interaction, which they adhere to, or vary pronoun uses if the referent has indicated such a preference.

$u_{d,i}$ Utterance including a pronominal reference to t.

P The speaker's general prior on pronoun production.

P^t_d The speaker's prior on t's pronouns at the time of interaction d. The support of P^t_d is not necessarily pointwise, and its support and distribution are subject to adjustments between different interactions, for instance in case of offline feedback about a pronoun use.

T_d Topic for interaction d.

pro^o_n Pronoun usages witnessed by the speaker at all times and for any referent.

These variables are plated across the set D of all discourses (spoken or written) where the speaker has referred to t, the set I of all interactions in said discourse, and the set N of all interactions witnessed at all by the Speaker.

A Bayesian approach captures the intuition that some individuals may have more rigid "priors" over pronouns for specific speakers, and therefore choose to override the referent's choice of pronouns. While this relative stubbornness is expected among individuals who adhere to gender binaries, it could also arise in individuals who are willing to expand their pronominal inventory but struggle to do so without significant exposure to more diverse pronoun usages.

Note that our models do not assume any reliance on external characteristics. While we generally disagree with the practice, a speaker's prior belief over pronoun distribution could be jointly determined by both linguistic experience as well as the co-occurrence of such characteristics in order to account for intentional or unintentional misgendering.

4 Probabilistic graphical model of community norms

Speakers do not obtain their linguistic knowledge from pure distributional statistics. Rather, their preferences are contextualized by interactions with others in their language communities and through interactions with individuals that may reinforce those community beliefs. In cases where a speaker belongs to a community with practices that either accept and embrace — or deny — the practice of naming oneself (Lind, 2023), speaker priors are expected to be sampled from the community prior over pronouns as well.

For example, queer and cis-binary communities display clear differences in linguistic preferences and consensus about whether one's pronouns neatly correspond to one's current presentation suggests (Rose et al., 2023). This gives rise to the prediction that some speakers will not readily adapt to signals that (in a given conversation) the relevant pronouns to use belong to some set and not others (Arnold et al., 2024), particularly if their linguistic knowledge strictly excludes gender neutral or neopronouns. On the other hand, queer folks who have many friends whose pronouns fall outside the gender binary can be expected to have more flexible and more uniform beliefs about potential pronouns.

At the scale of a whole community, where pronoun usage witnessed by someone are those produced by other members of a the community, our model becomes that of Figure 2: for all triplets c, s, t of individuals in a community C, $\text{pro}^{s,t}$ is a pronoun used by a s to refer to t and $P^{c,t}$ and P^c are the priors of c about possible pronoun usages, respectively for t specifically, and for anyone. Note that the self-referring case $s = t$ is not excluded, and is in fact an important part in building priors for the rest of the community.

5 Related work

A challenge for modeling pronoun use in practical systems arises when we presuppose that learning words boils down to the problem of mapping form onto meaning. For instance, early connectionist ap-

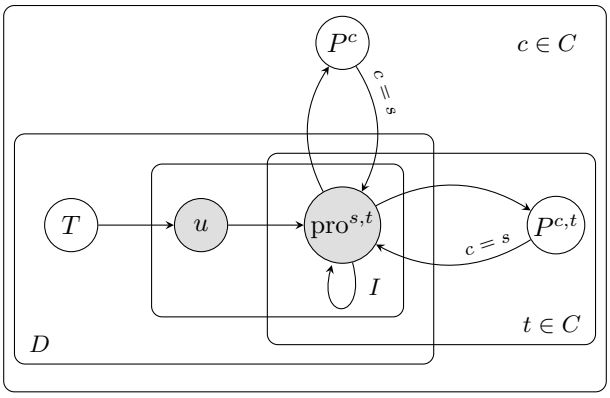

Figure 2: Community model with many speakers.

proaches to semantic representation, have treated the "meaning" of a word as a sparse d-dimensional vector consisting of several manually-selected semantic features (Cree et al., 2006; Rumelhart et al., 1986). Here, we propose that meaning be defined symbolically at the level of a referent rather than distributed across semantic features.

In word vectors trained on corpora, a "gender subspace" commonly emerges (Bolukbasi et al., 2016) that encodes social biases about canonical genders (e.g., stereotypes about the gender of nurses versus doctors). Pronouns and other high-frequency gendered nouns (e.g., man, woman) typically serve as critical anchors in the debiasing process, and serve as an excellent probe into the origins of biases in modern statistical NLP systems. Others have successfully demonstrated that non-binary pronoun LLM representations can be debiased, suggesting that the form-to-meaning mapping can be partially undone for novel referential forms (van Boven et al., 2024).

Being able to appropriately select the correct pronoun for a referent, as in text generation applications, is critical for ensuring equity and access to modern-day NLP tools. A number of studies have attempted to study gender bias in pronoun production. However, few of these studies have been able to probe the representations of pronouns, neopronouns, and name use that differs from the mainstream (Sakaguchi et al., 2021). The model we present here is capable of generating a wide variety of potential sentences to test the role of experience during fine-tuning of language models and thus improve gender inclusivity.

The present work is strongly informed by the integrative account presented in Ackerman (2019), who stated that cognitive, biological, and social factors combine to influence coreference resolution for non-binary people. They highlight that normatively unexpected mappings can nevertheless be made felicitous with sufficient supporting context.

6 Future Work

Our models allow for a straightforward integration of both witnessed pronoun uses and external priors in the process of pronoun selection in production. This provides a reasonably simple way to model pronoun acquisition during a long history of interactions in communities. However, for the sake of simplicity, certain interaction dynamics are not taken into account, and we leave to future work the search for improved models that balance the insights added by these refinements and the extra complexity that they would induce.

Our community model does not explicitly include non-linguistic social dynamics. Most importantly, language uses witnessed by a comprehender might have different weights depending on the speaker. For instance, the credit given to pronoun uses by speaker s for referent t could vary depending on how close to t s is assumed to be, and the $s = t$ case could be given a separate treatment. Furthermore, our models are only concerned with pronouns, which have the lightest semantic content of all referring expressions. However, it is likely that in practice, pronoun usage is also informed by the use of other referring expressions, such as names, formal titles, terms of address, etc. In our current model, these evidences are folded into the priors P, but more precise examination of their internal structure would provide a much richer model.

7 Conclusion

The model that we outline here shows that it is possible to represent individuals as possessing distributions of pronominal referring expressions, consistent with their own self-determined gender. The probabilistic graphical modeling accounts are flexible enough to allow learners to accommodate others based on their experience with linguistic variability in pronoun use. Additionally, the work provides a mechanism for the easy extension of one's linguistic vocabulary to incorporate novel pronouns, including but not limited to neopronouns, emojipronouns, and so on. We view this work as a critical bridge between cognitive scientific work on pronoun processing (Ackerman, 2019; Arnold et al., 2024; Rose et al., 2023) and computational modeling of linguistic variability (Eisenstein et al., 2010; Kleinschmidt et al., 2012) while also providing a way to advance equity in pronoun generation and comprehension (Ovalle et al., 2023; Piergentili et al., 2024; Lauscher et al., 2023).

References

Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4(1).

Amr Ahmed, Liangjie Hong, and Alexander Smola. 2013. Nested chinese restaurant franchise process: Applications to user tracking and document modeling. In *International Conference on Machine Learning*, pages 1426–1434. PMLR.

Jennifer E Arnold, Ranjani Venkatesh, Zachary Vig, Jennifer E Arnold, Ranjani Venkatesh, and Zachary A Vig. 2024. Gender competition in the production of nonbinary 'they'. *Glossa Psycholinguistics*, 3(1).

David Blei, Andrew Ng, and Michael Jordan. 2001. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482.

George S Cree, Chris McNorgan, and Ken McRae. 2006. Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(4):643.

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 Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA. Association for Computational Linguistics.

N.J. Enfield and T. Stivers. 2007. *Person Reference in Interaction: Linguistic, Cultural and Social Perspectives*. Language Culture and Cognition. Cambridge University Press.

Sourojit Ghosh and Aylin Caliskan. 2023. ChatGPT perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 901–912, New York, NY, USA. Association for Computing Machinery.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. 2023. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. *Advances in Neural Information Processing Systems*, 36:63687–63723.

Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2022. Uncertainty and inclusivity in gender bias annotation: An annotation taxonomy and annotated datasets of British English text. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 30–57, Seattle, Washington. Association for Computational Linguistics.

Dave F Kleinschmidt, Alex B Fine, and T Florian Jaeger. 2012. A belief-updating model of adaptation and cue combination in syntactic comprehension. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy. 2023. What about "em"? How Commercial Machine Translation Fails to Handle (Neo-)Pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–392, Toronto, Canada. Association for Computational Linguistics.

Miriam Lind. 2023. How to do gender with names: The name changes of trans individuals as performative speech acts. *Journal of Language and Sexuality*, 12(1):1–22.

Ashley R Moore, James Coda, Julia Donnelly Spiegelman, and Melisa Cahnmann-Taylor. 2024. Queer breaches and normative devices: language learners queering gender, sexuality, and the l2 classroom. *International Journal of Bilingual Education and Bilingualism*, 27(5):675–688.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1246–1266, New York, NY, USA. Association for Computing Machinery.

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2024. Enhancing gender-inclusive machine translation with neomorphemes and large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 300–314, Sheffield, UK. European Association for Machine Translation (EAMT).

Ell Rose, Max Winig, Jasper Nash, Kyra Roepke, and Kirby Conrod. 2023. Variation in acceptability of neologistic English pronouns. *Proceedings of the Linguistic Society of America*, 8(1):5526.

David E. Rumelhart, James L. McClelland, and PDP Research Group. 1986. *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press.

Keisuke Sakaguchi, Roland Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Transactions of the Association for Computing Machinery*, 64(9):99–106.

Eddie Ungless, Bjorn Ross, and Anne Lauscher. 2023. Stereotypes and smut: The (mis)representation of non-cisgender identities by text-to-image models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7919–7942, Toronto, Canada. Association for Computational Linguistics.

Goya van Boven, Yupei Du, and Dong Nguyen. 2024. Transforming Dutch: Debiasing Dutch Coreference Resolution Systems for Non-binary Pronouns. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 2470–2483, New York, NY, USA. Association for Computing Machinery.

Lal Zimman. 2019. Trans self-identification and the language of neoliberal selfhood: Agency, power, and the limits of monologic discourse. *International Journal of the Sociology of Language*, 2019(256):147–175.