Rethinking Identity in Natural Language Processing

Arianna Muti[1] **Elisa Bassignana**[2,3] **Debora Nozza**[1]
[1]Bocconi University, Italy
[2]IT University of Copenhagen, Denmark
[3]Pioneer Center for AI, Denmark
{arianna.muti, debora.nozza}@unibocconi.it elba@itu.dk

Abstract

Language is a primary site where identity is expressed and negotiated. Identity is widely understood as a relational and socially constructed phenomenon, encompassing both personal and social dimensions and emerging dynamically through language and social practice. Explicitly engaging with identity is essential for building NLP systems that are socially grounded and responsive to human language, yet NLP approaches often reduce it to static and easily measurable proxies, reducing its complexity to fixed categories or surface-level attributes. Through a systematic survey of ACL papers referencing "identity", we show that this reduction shapes how identity is operationalized across tasks. Limitations in how NLP engages with identity affect the whole NLP pipeline, including annotation, modeling and evaluation. We call for a shift toward more identity-aware language technologies.

Creative artifact: https://youtu.be/b80gr33w-70

1. Introduction

With the advent of generative AI, Natural Language Processing (NLP) systems increasingly interact with users in ways that require sensitivity to who people are, how they speak, and how meaning is shaped by social context. Yet, despite the centrality of identity to language use, much of NLP research treats identity in a narrow and reductive way. The dominant paradigm flattens identity into a limited set of demographic variables—such as gender, race, age, or nationality—assuming that these categories sufficiently capture meaningful variation in language. This flattening of identity affects core NLP tasks by obscuring annotators' interpretive frameworks and reducing personalization to coarse demographic proxies, often leading models to misinterpret intent and reinforce stereotypes.

This work examines how the concept of identity is (not) operationalized in NLP research. We first outline a theoretical definition of identity drawn from social and linguistic scholarship. We then conduct a systematic review of papers in the ACL Anthology that engage with the term "identity". Our analysis shows that existing work lacks a clear definition and instead treats identity as a proxy for demographics, user attributes, language varieties or chatbots/LLMs/agents. We conclude by proposing a roadmap for developing identity-aware NLP tasks, systems, and evaluation frameworks.

2. Defining Identity

Although there is no consensus on a single definition of identity, it is widely understood as a relational and socially constructed phenomenon. Social identity theory distinguishes between social and personal identity (Turner, 1987). In this framework, social identity refers to the aspects of the self that derive from membership in socially meaningful groups, while personal identity captures attributes that differentiate an individual from others. We operationalize social identity as group-level categories commonly encoded in datasets and annotations, including sociodemographic attributes (e.g., age, gender, race, ethnicity, nationality, disability status, religion, and sexual orientation) as well as affiliation-based categories (e.g., political alignment or other group memberships). Essentialist accounts of social identity assume clear group boundaries and internal homogeneity (Holliday, 1999, 2011; Bucholtz, 2003). In contrast, personal identity encompasses individual-level characteristics that are not reducible to group membership, such as personality traits, personal values, lived experiences, and skills. In this case, identity is seen as dynamic, context-dependent, and enacted through language, including participation in "small cultures" shaped by shared practices and linguistic repertoires (Holliday, 1999; Block, 2007).

3. Literature Review and Analysis

To understand how the concept of identity is operationalized within NLP research, we conduct a systematic literature review based on the PRISMA guidelines (Page et al., 2021). Our goal is twofold: (1) to identify which dimensions are treated as constitutive of identity in NLP papers and (2) to examine the contexts and tasks in which identity is invoked (e.g., dataset annotation, bias evaluation, hate speech detection, personalization, or model alignment). We used the ACL Anthology API to find all papers whose title or abstract contains the

keywords *identity* or *identities*. Our search included all work published before February 2026. We expand the pool of papers by adding twelve papers from the 2025 edition of the "Identity-aware AI and Awareness in Learning Agents Workshop".

One of the authors manually read the title and the abstract and filtered out papers that were not relevant to our definition of identity. Specifically, we discarded work where identity denoted: (i) identity relations in the context of coreference resolution and information extraction, (ii) authorship attribution (the "identity" of the author), (iii) named entity recognition (the "identity" of an entity), (iv) speaker identification in speech processing, and (v) privacy-preserving methods concerned with revealing or masking user identity. After filtering, our final corpus consists of 202 papers.

Our analysis suggests that NLP research operationalizes identity along four dimensions: as a fixed attribute of individuals (demographics), as behavioral patterns inferred from data (user), as linguistic indexicality (linguistic identity), or as a simulated property of systems (chatbots/LLMs/agents). Below, we explain and give examples for each dimension.

Demographics The largest proportion of papers falls under this category. In these works, identity is primarily operationalized as a set of demographic or group-based attributes. The categories we found in the papers include gender, race, ethnicity, nationality, political orientation, age, religion, sexual orientation, class, occupation, education, appearance, personality traits, disability, mental health conditions and astrological signs. In this category identity is treated as a proxy for socially salient group membership, particularly in contexts where such groups are associated with discrimination or marginalization. These identity categories are used in downstream tasks such as dataset annotation, fairness and bias evaluation, identifying whether content references or targets specific groups, and analyzing model behavior across identity mentions, especially in the context of harmful content. This framing aligns with social identity, in which social groups are treated as internally coherent and clearly bounded.
Identity: fixed static categories.

User In this category, identity refers to the behavioral patterns of users in online settings. Identity is operationalized through interaction histories, preferences, posting patterns, or engagement signals. Tasks include user profiling, personalization, recommendation systems, and the analysis of self-presentation in social media.
Identity: patterned behavior over time.

Linguistic identity In this category, identity is operationalized through patterns of language use that index membership in a social group. These works typically treat identity as something inferred from a speaker's linguistic repertoire—e.g., dialect features (such as African American English), code-switching practices, lexical choices (e.g., slang), and stylistic or pragmatic conventions. These features are often used as proxies for group membership, as, for example, community-specific slang can signal affiliation. At the same time, such features can be strategically used to obscure meaning for outsiders. A salient case is the use of dogwhistles: terms or expressions that appear innocuous in general usage but carry derogatory meanings within a particular community. **Identity:** linguistic repertoire.

Chatbots/LLMs In this category, identity refers to artificial agents and the personas they adopt or are assigned. Papers explore how conversational agents construct, simulate, or are perceived as having identities, including demographic traits, personality characteristics, or social roles. Identity is therefore conceptualized both as a design feature (e.g., persona-based dialogue systems) and as an emergent property of model behavior.
Identity: constructed persona of an artificial agent.

3.1. Current Limitations and Proposed Solutions

Our analysis highlights several limitations in how identity is currently treated in NLP:

Conceptual Ambiguity The term *identity* is used inconsistently, often interchangeably with demographics, user profiling, or protected attributes. This lack of conceptual clarity weakens theoretical grounding and risks oversimplification.
⇒ Use precise terminology when appropriate and reserve the term *identity* for cases where it is explicitly defined and theoretically grounded.

Reductionism Identity is frequently reduced to binary or coarse-grained categories (e.g., male/female, liberal/conservative), ignoring fluidity and thereby oversimplifying social reality. This reduction is often driven by the need to fit identity into discrete labels for supervised classification. Beyond limiting granularity, NLP work also tends to operationalize identity as something that can be externally inferred from observable signals (e.g., text or metadata), privileging socially ascribed categories over individuals' self-understanding. In doing so, it foregrounds social identity, while overlooking personal identity. With LLMs, however, both inputs and outputs can accommodate richer, self-described,

and context-sensitive identity representations. ⇒ Move beyond fixed, externally imposed categories by incorporating self-identification and allowing for richer, context-sensitive representations of identity.

Static Modeling Most NLP systems treat identity as a fixed label attached to users or texts. However, identity is relational and interactional: meaning emerges through the interplay of social factors such as speaker and receiver characteristics, social relations, context, social norms, culture and ideology, and communicative goals (Hovy and Yang, 2021). Treating identity as a static variable collapses these dimensions into a single feature and ignores how identity shifts across audiences, situations, and time. ⇒ Move from static, speaker-based representations to contextualized modeling frameworks that incorporate interactional variables and longitudinal analysis.

Evaluation We propose evaluation tasks for identity-aware NLP focused on three areas: (1) **Subjective tasks**, assessing whether LLMs account for identity-related constraints rather than assuming a default user; (2) **Refusals**, auditing whether similar requests receive asymmetric responses depending on identity framing; and (3) **Identity representation in training data**, examining whether models reflect diverse identities or disproportionately align with WEIRD (Western, Educated, Industrialized, Rich, Democratic) norms (Santy et al., 2023).

Acknowledgements

We thank the MilaNLP group at Bocconi University for feedback. Arianna Muti and Debora Nozza are supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE). Arianna Muti and Debora Nozza are members of the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis. Elisa Bassignana is supported by a research grant (VIL59826) from VILLUM FONDEN.

References

David Block. 2007. *Second Language Identities*. Continuum, London, England.

Mary Bucholtz. 2003. Sociolinguistic nostalgia and the authentication of identity. University of California.

Adrian Holliday. 1999. Small cultures. *Applied Linguistics*, 20(2):237–264.

Adrian Holliday. 2011. *Intercultural Communication and Ideology*. Sage, London.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372(n71). Published 29 March 2021, open access.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

John Charles Turner. 1987. *Rediscovering the Social Group: A Self-Categorization Theory*. Basil Blackwell, Oxford.