Implicit Cultural Identity Signals in Language: Detection and Effects in Negotiation Dialogue

Bin Han, Danah Yun, James Hale, Jonathan Gratch
University of Southern California
Los Angeles, CA, USA
{binhan, dnyun, jahale}@usc.edu, gratch@ict.usc.edu

Abstract

Language transmits cultural identity even without explicit disclosure, shaping how individuals perceive and engage in interpersonal tasks. We investigate these identity signals within the KODIS (KObe DISpute) corpus, an English-only corpus of anonymous text-chat negotiations involving participants from the US, UK, Mexico, and South Africa. We test whether a speaker's country can be inferred from dialogue using zero-shot LLMs and embedding-based classifiers. Our results demonstrate that cultural identity is reliably detectable in text, with embedding models achieving far higher accuracy. Critically, while objective negotiation outcomes remain consistent across groups, these subtle identity cues significantly alter participants' subjective feelings about the interaction. These findings suggest that cultural identity-related signals are embedded in language and may be relevant for analyzing negotiation dialogue.

Keywords: Cultural Identity, Sociolinguistic Cues, Social Interaction, Person Perception

1. Introduction

Language use can convey cues about a speaker's background and social identity (Gumperz, 1982). Word choice and expression shape impressions about a person's background and social identity, and these impressions influence how interaction unfolds (Koschate et al., 2021). In social psychology, this process is known as *Person Perception* (Keith, 2013). Interaction is influenced not only by what is said but also by how the speaker is perceived (Kunda and Thagard, 1996). In natural dialogue, speakers rarely state their cultural background directly. Even so, patterns of language use may convey subtle identity-related signals.

Identifying these signals from text is a meaningful research goal, not just a step toward downstream analysis. In many natural dialogue datasets, especially those from online disputes or archived negotiations, cultural background information is missing or unreliable. A classifier that infers cultural origin from language makes culture observable and measurable in large unlabeled corpora where metadata is not available (Tetreault et al., 2013). Negotiation is a useful context to study this problem. Prior dialect identification work mostly focuses on neutral or everyday language, but negotiation dialogue is different. It involves strategic interaction, power differences, and emotionally charged face threats (Lawler and Yoon, 1995). These conditions may strengthen or weaken cultural signals, but this has not been tested empirically.

This paper examines whether traces of cultural identity emerge in task-based dialogue when speakers do not explicitly disclose their background. It also examines whether these signals reflect surface-level differences in expression or are associated with differences in interaction outcomes. To explore this, we analyze English negotiation dialogues collected under the same task conditions from participants in four countries (Hale et al., 2025). The analysis focuses on whether a speaker's country can be inferred from dialogue alone and whether the strength of this inference is related to negotiation behavior.

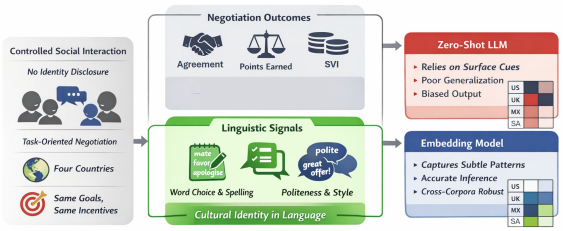

Figure 1: Overview of the study design and evaluation framework.

Specifically, we address the following research questions:

- **RQ1:** To what extent can a speaker's country of origin be inferred from dialogue when cultural identity is neither relevant nor intentionally disclosed?
- **RQ2:** Which computational approach better captures these signals in negotiation dialogue: zero-shot LLMs or embedding-based classifiers?
- **RQ3:** Are these inferred identity signals associated with differences in negotiation outcomes, as reflected in within-country versus between-country interactions?

Proceedings of the 1st SoCon Workshop & the 2nd NLPSI Workshop @ LREC 2026, pages 34–39
12 May 2026.

2. Related work

2.1. Cultural identity in language

Sociolinguistic research has shown that social and cultural identity can be reflected in patterns of language use, such as lexical choice, spelling conventions, and pragmatic style (Bucholtz and Hall, 2005; Labov, 2006). These patterns often emerge without explicit self-disclosure and can be observed across different communicative contexts (Bucholtz and Hall, 2005). Prior work suggests that listeners may form social impressions based on such linguistic cues, even when identity is not directly relevant to the task (Giles and Powesland, 1975).

2.2. NLP for social variable detection

Prior work in NLP has examined the prediction of social variables from text, including demographic and cultural attributes (Tetreault et al., 2013). These studies are motivated by the observation that such social variables are systematically reflected in linguistic patterns, and have proposed a range of computational approaches, most commonly using feature-based representations of text (Zampieri et al., 2018), with more recent work applying large language models for social attribute prediction (Al-Nuaimi et al., 2024). However, these studies have largely focused on static text genres such as essays or social media posts, overlooking interactive task-oriented dialogue. Negotiation is a particularly compelling yet underexplored setting: it is fundamentally a social interaction shaped by perceptions of group membership and identity (Tajfel et al., 2001), where in-group favoritism and intergroup bias can influence communication style and subjective evaluations (Brewer, 1999; De Dreu and Carnevale, 2003). However, the application of NLP methods to negotiation dialogue remains largely unexplored, even though strategic pressure and emotional intensity may influence how cultural identity is expressed.

3. Method

3.1. KODIS Dataset

We use English negotiation dialogues from the KODIS dataset. The dialogues are collected from an online negotiation experiment in which participants interact anonymously to resolve a purchase dispute (Hale et al., 2025). Participants do not know each other's country of origin, and there is no requirement or incentive to reveal it during the interaction. All dialogues follow the same task scenario: a buyer requests a refund after receiving an incorrect basketball jersey from a seller. The scenario is designed to elicit emotionally charged exchanges, as the buyer has been wronged and may express frustration or anger during the dispute.

Each dialogue is associated with a buyer country label among four classes: *U.S., U.K., Mexico, and South Africa*.We construct a balanced dataset with 80 dialogues per country (320 total), where both buyer and seller are from the same country.

3.2. Models

We analyze only the buyer side of each dialogue (speaker-specific setting). We use 5-fold stratified cross-validation with an 80% train and 20% test split in each fold.

- **LLM (Zero-shot).** We use `gpt-4o-mini` model as a zero-shot classifier. The prompt provides a short list of linguistically grounded cues (e.g., spelling variants, lexical choice, delivery terminology, and interactional tone) to guide the prediction. The model is queried with a temperature of 0.3 to reduce output variability. [1]

- **Embedding + Logistic Regression.** Buyer utterances are embedded using `text-embedding-3-large` (3072-dimensional representations). A multinomial logistic regression classifier is trained on these embeddings using L_2 regularization with a maximum of 1000 iterations. The random seed is fixed for reproducibility.

- **Embedding + SVM.** The same precomputed embeddings are used as input features. We train a Support Vector Machine with an RBF kernel, using the library's standard regularization and kernel parameters.

- **Embedding + Random Forest.** The same embedding representations are used to train a Random Forest classifier. We use the default number of trees and feature-sampling strategy provided by the library.

‘

3.3. Lexical Cues by Country

We analyze TF–IDF logistic regression coefficients to identify lexical features associated with each country. Features are ranked by their class-specific coefficient magnitude, and the top positively weighted cues are selected for inspection. The model assigns high weights to words and phrases that reflect country-specific linguistic style.

[1] The full prompt is provided in the Appendix.

3.4. Within- vs. Between-Country Interaction Analysis

We compared negotiation outcomes between within-country and between-country dyads. All dialogues were collected under identical task conditions with matched incentives and roles. We examined resolution rates, buyer and seller points, joint points, and all facets of the Subjective Value Inventory (SVI) (Curhan et al., 2006). For this analysis, we included all dyads with complete outcome information ($N = 2076$ dialogues). Dyads were categorized as within-country or between-country based on whether the buyer and seller shared the same country label. Resolution outcomes were analyzed using a chi-square test of independence, while buyer points, seller points, joint points, and SVI measures were analyzed using mixed-effects ANOVAs with Role as a within-dyad factor and Match as a between-dyad factor.

4. Result

4.1. Performance

Table 1: Model performance comparison.

Method	Accuracy	F1-Score
LLM	0.431	0.374
LogReg	**0.625**	**0.627**
SVM	0.616	0.628
RandomForest	0.615	0.612
Chance (4-class)	0.250	–

In Table 1, embedding-based methods substantially outperform the LLM classifier across all metrics. Logistic Regression achieves 62.5% accuracy and an F1-score of 0.627, compared to 43.1% accuracy and 0.370 F1 for the LLM.

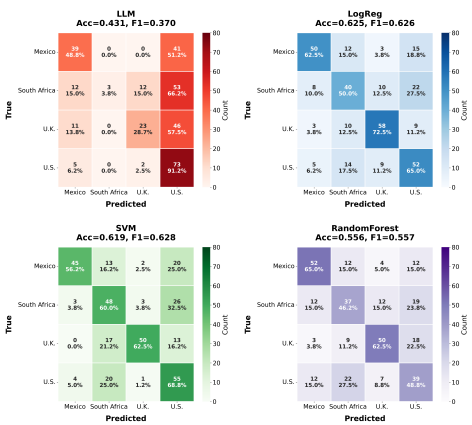

Figure 2: Confusion matrices for four models on the four-country classification task

The confusion matrices in Figure 2 further clarify why this gap emerges. The LLM exhibits a strong prediction bias toward the U.S. class: for true U.S. instances, it predicts U.S. 91.2% of the time, but it also incorrectly maps a large proportion of Mexico (51.2%), South Africa (66.2%), and U.K. (57.5%) samples into the U.S. category.

In contrast, embedding-based classifiers show balanced decision boundaries across all four classes. Logistic Regression correctly identifies 62.5% of Mexico, 50.0% of South Africa, 72.5% of the U.K., and 65.0% of U.S. samples, with substantially reduced cross-country confusion. This balanced error distribution leads to a much higher macro F1-score. Similar patterns are observed for SVM (F1=0.628) and Random Forest (F1=0.557), all of which avoid the systematic overprediction seen in the LLM.

Limitations of Zero-Shot LLM Classification: Analysis of the explanations ("reasons") generated by the LLM alongside its predictions reveals two main sources of error. First, the model shows a strong bias toward the U.S. class, frequently interpreting generic negotiation language (e.g., references to refunds or shipping) as evidence of American origin, regardless of the speaker's actual country. Second, the LLM relies primarily on explicit surface-level cues, such as spelling variants or named institutions, and fails to capture more subtle discourse-level differences when these markers are absent. These tendencies are reflected in the model's own reasons, which often cite generic lexical cues rather than distinctive linguistic patterns.

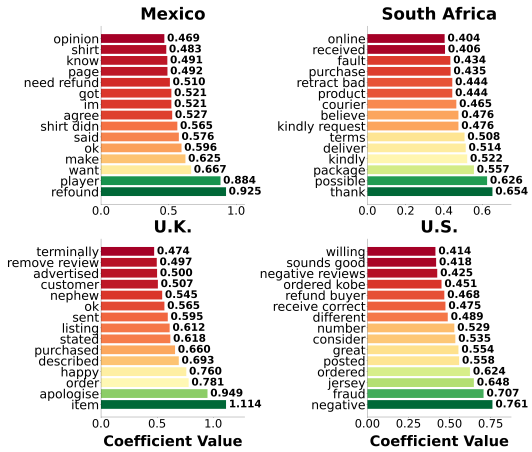

Figure 3: Lexical features with the highest coefficients for each country in the logistic regression model.

4.2. Corpus-level Linguistic Signals

Figure 3 shows distinctive words for each country, identified based on the highest positive logistic regression coefficients from the country-level classifier. These features reflect corpus-level stylistic tendencies rather than fixed cultural traits, and may

be shaped by task context, role expectations, and L2 language use. U.K. dialogues contain politeness markers and British spelling variants (e.g., “apologise”, “colour”). U.S. dialogues emphasize direct and transactional language (e.g., “refund”, “fraud”). Mexican dialogues show simplified phrasing and L2-influenced constructions (e.g., “im”, “ok”). South African dialogues contain procedural and regulatory terms (e.g., “courier”, “terms”).

4.3. Within- vs. Between-Country

Table 2: Result of Mixed ANOVA Analysis

DV	Effect	p	Direction
Points	Role	<0.001	Buyer $>$ Seller
SVI Instrumental	Match	.040	Between $>$ Within
SVI Instrumental	Role	.022	Buyer $>$ Seller
SVI Relationship	Match	.004	Between $>$ Within
SVI Relationship	Role	.017	Buyer $>$ Seller

We analyzed interaction outcomes using mixed ANOVAs with Match (within-country vs. between-country) as a between-dyads factor and Role (buyer vs. seller) as a within-dyads factor. Resolution outcomes were analyzed separately using a chi-square test. Table 2 summarizes the main effects of Match and Role on negotiation outcome measures. Significant effects were observed for Role on points and for Match on instrumental and relationship-focused subjective value.

5. Discussion

Regarding RQ1, our results suggest that cultural identity-related signals may emerge through linguistic style even when such information is not explicitly disclosed. The results support this assumption: despite the absence of direct references to nationality, both zero-shot LLMs and embedding-based models were able to infer speakers’ countries from dialogue patterns alone. This suggests that social identity may be reflected not only in explicit content but also in subtle stylistic features such as orthographic choices, hedging, and discourse structure.

For RQ2, embedding-based classifiers captured identity signals more robustly than the zero-shot LLM in this setting. Although zero-shot LLMs are often effective in many NLP settings, they were less robust than embedding-based classifiers in our specific negotiation classification setup. One possible reason is that negotiation dialogue contains fewer explicit nationality markers than more conventional dialect identification settings Error analysis further indicated that the LLM often relied on a small number of salient lexical cues and, when such cues were weak or absent, tended to default to predicting the U.S. class. This observation suggests that, in our setting, the zero-shot model was more sensitive to sparse surface cues, while embedding-based models may have captured more distributed stylistic patterns.

For RQ3, the comparison between within-country and between-country interactions revealed an interesting dissociation: despite the competitive and emotionally charged nature of the negotiation task, objective outcomes did not significantly differ across conditions, whereas some subjective evaluations did. A more conservative interpretation is that identity-related cues may shape how interactions are subjectively experienced, even when they do not translate into measurable differences in objective outcomes. In such contexts, even emotionally intense exchanges may be interpreted more charitably in cross-country interactions, shaping relational perceptions without altering material negotiation outcomes. Together, these findings suggest that social signals expressed through language may influence how interactions are experienced and evaluated, rather than the objective results of the negotiation itself.

6. Conclusion

Overall, our findings show that implicit cultural identity cues can be detected in negotiation language. Embedding-based classifiers consistently outperformed zero-shot LLM predictions, suggesting that distributed stylistic patterns may provide more reliable signals for identity inference than surface-level reasoning. These results suggest that culture classification can serve as a useful measurement step for making country-linked variation observable in naturalistic dialogue data, enabling the study of cultural effects even when explicit metadata is unavailable. They also provide a starting point for examining when cultural signals become salient or are overridden in contexts shaped by strategic interaction and emotional intensity.

Several limitations should be noted. First, participants were not directly asked about their partner’s nationality or identity perception, making it difficult to determine whether identity inferences were consciously recognized or only implicitly processed. Second, although we compared multiple classifiers and a zero-shot LLM, additional model families or fine-tuned approaches may yield different patterns of robustness and generalization. Future work could examine how emotional dynamics interact with identity signaling, particularly by considering the role of tone, politeness, and attribution in text-based communication. It would also be valuable to investigate whether the strength or accuracy of inferred identity signals acts as a moderator that shapes partner responses and the unfolding of interaction dynamics.

Acknowledgements

Research was sponsored by the Air Force Office of Scientific Research under grant FA9550-23-1-0320. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

Khaled AlNuaimi, Gautier Marti, Mathieu Ravaut, Abdulla AlKetbi, Andreas Henschel, and Raed Jaradat. 2024. Enriching datasets with demographics through large language models: What's in a name? *arXiv preprint arXiv:2409.11491*.

Marilynn B Brewer. 1999. The psychology of prejudice: Ingroup love and outgroup hate? *Journal of social issues*, 55(3):429–444.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.

Jared R Curhan, Hillary Anger Elfenbein, and Heng Xu. 2006. What do people value when they negotiate? mapping the domain of subjective value in negotiation. *Journal of personality and social psychology*, 91(3):493.

Carsten KW De Dreu and Peter J Carnevale. 2003. Motivational bases of information processing and strategy in conflict and negotiation.

Howard Giles and Peter F Powesland. 1975. *Speech style and social evaluation.* Academic Press.

John J Gumperz. 1982. *Language and social identity*. 2. Cambridge University Press.

James Anthony Hale, Sushrita Rakshit, Kushal Chawla, Jeanne M Brett, and Jonathan Gratch. 2025. Kodis: A multicultural dispute resolution dialogue corpus. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12771–12785.

Kenneth D. Keith. 2013. *Person Perception*, pages 991–993. John Wiley & Sons, Ltd.

Miriam Koschate, Elahe Naserian, Luke Dickens, Avelie Stuart, Alessandra Russo, and Mark Levine. 2021. Asia: Automated social identity assessment using linguistic style. *Behavior Research Methods*, 53(4):1762–1781.

Z. Kunda and P. Thagard. 1996. Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, 103:284–308.

William Labov. 2006. *The social stratification of English in New York city*. Cambridge University Press.

Edward J Lawler and Jeongkoo Yoon. 1995. Structural power and emotional processes in negotiation: A social exchange approach.

Henri Tajfel, John Turner, William G Austin, and Stephen Worchel. 2001. An integrative theory of intergroup conflict. *Intergroup relations: Essential readings*, pages 94–109.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 48–57.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Dirk Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 1–17, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Appendices

A. LLM Prompt

LLM Prompt Snippet

System:
You are a helpful assistant that identifies the country of origin from a dialog.

User:
Below is a dialog from a customer service dispute over the purchase of an American basketball player jersey. Both buyer and seller are from the same country: U.S., U.K., Mexico, or South Africa.

Use the following linguistic clues to help decide:

- Spelling: "colour/organisation" (U.K.) vs "color/organization" (U.S.)

- Word choice: “advert” (U.K.) vs “ad” (U.S.)
- Spanish L2 typos: “recieved”, “missunderstanding” (Mexico)
- Delivery terms: Royal mail (U.K.), shipping (U.S.), courier (South Africa), parcel service (Mexico)
- Tone and references: PROFECO (Mexico), Consumer Protection Act (South Africa)
- Ignore currency formatting and the fact that the item is U.S. merchandise

Think step by step using these clues, but do not show your reasoning.

Dialog:
<*Speaker text inserted here*>

Output format:
Prediction: `<country>`, Reason: `<reason>`