



Artificial Intelligence-Based Corpus Analysis of Linguistic Identity in Contemporary Pakistani Fiction

Taqdees Mansoor Abbasi^a, Muhammad Aslam Khan^b, Sohail Falak Sher^c

- a. M.Phil Scholar, Department of English Literature, The Islamia University of Bahawalpur
- b. PhD Scholar in the Department of English Linguistics, The Islamia University of Bahawalpur, Pakistan
- c. Assistant Professor, Department of English Literature, The Islamia University of Bahawalpur

ARTICLE INFO

Received:

July 17, 2026

Revision Received:

August 04, 2026

Accepted:

August 25, 2026

Available Online:

September 11, 2026

Keywords:

linguistic identity, corpus stylistics, Pakistani English, contemporary Pakistani fiction, code-switching, natural language processing, topic modelling, and postcolonial literature

ABSTRACT

This study examines how linguistic identity is constructed in contemporary Pakistani fiction written in English, using an artificial intelligence (AI)-assisted corpus-analytic framework. A purpose-built corpus of 24 novels and short-story collections published between 2000 and 2023 (approximately 2.94 million running words) was compiled, digitized, and annotated. The corpus was analyzed through an integrated pipeline that combined classical corpus-linguistic techniques (keyword extraction, collocation, concordance, n-gram profiling) with contemporary natural language processing methods (transformer-based part-of-speech tagging and named entity recognition, language identification for code-switching detection, sentence-embedding clustering, BERTopic modelling, and aspect-level sentiment and stance classification). Findings indicate that linguistic identity in the corpus is realized through five interlocking systems: (i) dense lexical borrowing from Urdu, Punjabi, Sindhi, and Pashto, concentrated in the semantic fields of kinship, food, religion, clothing, and address; (ii) intra-sentential code-switching that functions as an index of intimacy, authenticity, and social positioning rather than as a communicative deficit; (iii) a distinctive collocational grammar around identity nouns such as 'home', 'country', 'people', and 'name'; (iv) pronominal patterning in which first-person plural forms carry the burden of collective and national self-definition; and (v) evaluative and stance patterns that are markedly ambivalent, combining affiliation with critique. Topic modelling recovered nine stable semantic themes, of which migration and belonging, religion and moral order, and urban violence were the most productive of identity language. The study argues that AI-based corpus analysis makes visible the systematic, distributional character of literary identity work that close reading can only sample, and it proposes an integrated AI-corpus model of linguistic identity for Anglophone postcolonial fiction. Methodological limitations, validation procedures, and directions for future research are discussed.

1. Introduction

Language is never a neutral instrument of storytelling. In multilingual societies it is one of the primary media through which writers position characters socially, historically and politically. Pakistan offers a particularly dense linguistic ecology for such positioning: Urdu operates as the national language and a symbol of post-Partition cohesion; Punjabi, Sindhi, Pashto, Balochi, Saraiki, and Brahui function as regional and affective mother tongues; Arabic and Persian supply a religious and literary

* Correspondence to: M.Phil Scholar, Department of English Literature, The Islamia University of Bahawalpur
E-mail address: taqdeesabbasi92@gmail.com (T.M. Abbasi)



register; and English occupies the contested position of an inherited colonial language that has become an internal marker of class, mobility, and access to global publics. Contemporary Pakistani fiction written in English is produced at the point where all these systems intersect, and its language therefore carries an unusually heavy indexical load. Anglophone Pakistani fiction has moved from the margins of South Asian writing to a position of international visibility. Novels by Mohsin Hamid, Kamila Shamsie, Mohammed Hanif, Nadeem Aslam, Uzma Aslam Khan, Bina Shah, Fatima Bhutto, H. M. Naqvi, and others circulate simultaneously in Karachi, London, and New York. This double address creates a linguistic problem that is also an identity problem: how does a writer render a Lahore courtyard, a Karachi newsroom, or a Peshawar bazaar in English without either flattening the local into transparency or exoticizing it for outsiders? The solutions writers adopt, including selective borrowing, untranslated insertion, glossing, calquing, syntactic transfer, and code-switching, constitute a repertoire of identity-signalling devices that can be observed, counted, and compared. Traditionally, such devices have been studied through close reading of a small number of canonical texts. Close reading remains indispensable for interpretation, but it is poorly suited to questions of distribution, frequency, and systematicity. It cannot easily establish whether a device is idiosyncratic to one author or characteristic of a literary formation, nor whether a pattern intensifies across a period. Corpus linguistics was developed to answer exactly such questions, and the recent maturation of AI-based language technologies has considerably enlarged what corpus analysis can detect. The present study brings these resources to bear on the linguistic construction of identity in contemporary Pakistani fiction.

Linguistic identity, in the sense adopted here, refers to the ways in which speakers and writers use linguistic resources to claim, ascribe, contest, or negotiate social belonging. Identity in this view is not a fixed attribute that language merely reports; it is an interactional and textual accomplishment produced through repeated acts of positioning. In fiction, this accomplishment is doubled: characters position themselves within the storyworld, and the narrating text positions the storyworld in relation to its readers. Pakistani fiction in English performs this double positioning in recognizable ways. Kinship terms such as *ammi*, *abba*, *khala*, *chacha*, and *beta* are retained in the original even when the surrounding narration is entirely English, so that intimacy is marked as untranslatable. Religious formulae such as *inshallah*, *mashallah*, *alhamdulillah*, and *astaghfirullah* are used less as doctrinal statements than as conversational punctuation that indexes a shared moral universe. Food, dress, and domestic architecture supply a further lexical layer, with items such as *biryani*, *paratha*, *dupatta*, *shalwar kameez*, *charpai*, and *mohalla* appearing without italics in many recent texts, a typographic decision that itself asserts normality rather than exoticism. Beyond the lexicon, syntactic and pragmatic transfer from Urdu, such as tag-question patterns, reduplication, and the use of 'na' and 'yaar' as discourse particles, produces a recognizable literary Pakistani English. These features are not decorative. They construct positions: the diasporic returnee whose Urdu has become effortful; the elite Karachi speaker whose English is unmarked and whose Urdu is performed; the servant, the driver, or the tenant farmer whose speech is rendered in a translated register that signals class distance; the militant or the cleric whose Arabic-heavy idiom marks ideological affiliation. A systematic account of linguistic identity in this literature must therefore describe not only which forms occur but also who uses them, in what company, and with what evaluative coloring.

Corpus stylistics established that literary meaning has a distributional dimension: recurrent collocations, keyword profiles, and clusters reveal patterns of characterization and worldview that are invisible in single passages. Until recently, however, corpus tools operated largely at the level of orthographic form. They could count 'home' but could not distinguish its senses; they could retrieve 'Ammi' but could not label it as an Urdu kinship term; they could not group thematically related passages that shared no vocabulary. AI-based methods have changed this. Transformer language models produce contextual representations in which the same word form receives different vectors in different environments, enabling sense-sensitive analysis. Sentence-embedding models allow passages to be clustered by meaning rather than by shared word forms. Neural sequence labellers permit reliable part-of-speech tagging and named entity recognition on non-standard, code-mixed text. Character-level language identification can flag embedded Urdu, Punjabi, or Arabic material even when it is Romanized inconsistently. Neural topic models, such as BERTopic, combine embeddings with clustering and class-based term weighting to produce topic solutions that are more coherent than purely count-based alternatives. Aspect-based sentiment and stance classifiers make it possible to ask not merely whether a passage is positive or negative, but how a particular identity category is evaluated within it. At the same time, these methods introduce risks. Models trained predominantly on Standard English and non-literary genres may mis-tag transliterated material, reproduce cultural bias in sentiment assignment, and produce topic solutions that are statistically stable but interpretively empty. AI-based literary analysis therefore requires an explicit validation architecture in which machine output is checked against human annotation and concordance evidence. The present study treats such validation as a constitutive part of the method rather than an afterthought.

1.1. Research Problem

Three gaps motivate this study. First, existing scholarship on Pakistani Anglophone fiction is overwhelmingly interpretive and text-selective. Valuable readings exist of individual novels, but there is little evidence about whether the linguistic strategies identified in those readings are shared across the field, and if so, in what proportions. Second, corpus-based work on South Asian English has concentrated on journalistic, academic, and conversational registers; literary corpora of Pakistani English

remain small, unstandardized, and rarely publicly documented. Third, where computational methods have been applied to postcolonial fiction, they have generally been confined to frequency counts and simple keyword lists, without the semantic layer that contemporary AI methods make available. Therefore, the problem addressed here can be stated as follows: there is no systematic, replicable, semantically informed account of how linguistic identity is constructed across contemporary Pakistani fiction as a body of writing rather than as a set of individual texts. Without such an account, claims about hybridity, code-switching, and the venularization of English in this literature rest on illustrative rather than distributional evidence.

1.2. Aim and Research Questions of the Study

This study aims to develop and apply an AI-assisted corpus-analytic framework capable of describing the linguistic construction of identity in contemporary Pakistani fiction in English. The specific research questions are:

- **RQ1.** Which lexical and phraseological features most strongly characterize contemporary Pakistani fiction in English relative to a general reference corpus of contemporary fiction?
- **RQ2.** How is multilingual material from Urdu and regional languages distributed across semantic fields, speakers, and narrative levels, and what functions does it perform?
- **RQ3.** What collocational and semantic profiles attach to core identity terms such as home, country, people, name, language, and blood?
- **RQ4.** Which semantic themes does AI-based topic modelling recover, and how do these themes relate to identity construction?
- **RQ5.** How do stance and evaluation patterns across identity categories of gender, class, region, nation, and diaspora?
- **RQ6.** What does an AI-based corpus approach add to and what does it omit from interpretive accounts of the same literature?

2. Review of Related Literature

2.1. Language, Literature and Identity

Sociolinguistic theory has moved decisively away from treating identity as a demographic given. Identity is now widely understood as emergent in interaction, relational rather than absolute, partly deliberate and partly habitual, and always positioned with respect to available social categories. Literary discourse is an especially productive site for observing this because writers deliberately manipulate the indexical value of forms, assigning particular registers, borrowings, and speech habits to particular character positions. What in conversation is largely automatically becomes a designed system in fiction.

2.2. Pakistani English as a Variety

Descriptive work on Pakistani English has documented lexical borrowing, semantic shift, distinctive prepositional and article usage, and the incorporation of Islamic religious vocabulary into everyday English. Scholarship has also emphasized that Pakistani English is not a single object but a cline of varieties differentiated by education, region, and domain. Literary Pakistani English draws selectively on this cline: writers may reproduce acrolectal usage in narration while assigning mesolectal or basilectal features to particular characters, so that variation within the variety becomes a resource for characterization.

2.3. Language Politics and the Pakistani Linguistic Ecology

Research on language policy in Pakistan has documented persistent tension between the symbolic promotion of Urdu, the institutional dominance of English in elite education and administration, and the marginalization of regional languages in formal domains. These tensions supply much of the raw material for the identity conflicts staged in fiction. Characters who move between English-medium schooling and Urdu- or Punjabi-speaking households enact in miniature the stratification described in the policy literature, and the fiction frequently thematises the shame, advantage, and estrangement that follow from it.

2.4. Postcolonial Writing and Linguistic Appropriation

Postcolonial criticism has long theorized the strategies by which writers appropriate a colonial language: abrogation of the standard, untranslated insertion, glossing, interlanguage, and syntactic fusion. These strategies are directly observable in corpus data as ratios of borrowed to English lexis, presence or absence of in-text glossing, and typographic marking. Corpus analysis thus offers a way of testing critical claims about appropriation quantitatively rather than illustratively.

2.5. Code-Switching: Structural and Pragmatic Accounts

Structural accounts distinguish intersentential from intrasentential switching and specify constraints on where switches may occur, typically identifying a matrix language that supplies morphosyntactic frames and an embedded language that supplies content items. Pragmatic accounts emphasize the contextualizing function of switching: a switch signals a change of footing, addressee, key, or affiliation. In written fiction, switching is stylized rather than spontaneous, and therefore tends to be functionally concentrated in dialogue, reported thought, and culturally saturated noun phrases.

2.6. Corpus Linguistics and Corpus Stylistics

Corpus stylistics has demonstrated that literary texts can be described through the same distributional apparatus as other discourses: keyword analysis identifies lexis that is statistically over-represented relative to a reference corpus; collocation reveals habitual company; clusters and n-grams expose formulaic patterning; and concordance display returns the analyst to context. The tradition insists that quantitative salience is a prompt for interpretation, not a substitute for it, a principle retained in the present design.

2.7. Natural Language Processing for Literary Text

Computational literary studies have applied named entity recognition to character networks, sequence labelling to speech and thought representation, and embeddings to genre and style classification. Literary text poses specific difficulties for standard NLP: non-standard orthography, dense figurative language, long-range coreference, and, in multilingual literature, embedded material that falls outside the model's training distribution. Domain adaptation and manual correction of a sample are therefore standard requirements.

2.8. Topic Modelling and Semantic Clustering

Probabilistic topic models represent documents as mixtures over latent topics, and topics as distributions over words. Their weakness for literary corpora is sensitivity to preprocessing choices and a tendency to recover surface vocabulary clusters. Embedding-based topic modelling reduces this weakness by clustering dense sentence representations and then extracting class-distinctive terms, which typically yields more interpretable topics on smaller and stylistically varied corpora.

2.9. Sentiment, Stance and Appraisal

Sentiment analysis assigns valence, while stance analysis, drawing on appraisal frameworks, distinguishes affect, judgement of behavior, and appreciation of things. For identity research, this distinction matters because a passage may express negative affect about a place while expressing positive judgement of the people in it. Aspect-based methods, which attach polarity to specific targets rather than to whole sentences, are therefore preferred here.

2.10. Research Gap

Taken together, the literature establishes the sociolinguistic importance of identity work, the descriptive profile of Pakistani English, the critical vocabulary of appropriation, and the technical means of large-scale textual analysis. However, it does not yet provide their integration: a study that treats a defined body of contemporary Pakistani fiction as a corpus, analyses it with semantically informed AI tools, validates the output, and reads the results back into literary and sociolinguistic theory. The present study addresses this gap.

2.11. Theoretical Framework

This study adopts a four-layer framework that combines sociolinguistic, corpus-linguistic, and postcolonial premises. The first layer is indexicality and positioning. Linguistic forms acquire social meaning through repeated association with categories of person and situation; identity is achieved when a text deploys these indexical resources to occupy or assign a position. This layer motivates the analysis of borrowings, address terms, and switching as positioning devices rather than as ornament. The second layer is the distributional principle. Meaning is a function of company: a word is characterized by the collocates and colligations that recur with it. This layer motivates keyword, collocation, and n-gram analysis, and licenses the claim that identity has a measurable phraseological signature. The third layer is appropriation and hybridity. The postcolonial text works on the inherited language, marking distance from the metropolitan standard through insertion, calque, and refusal to gloss. This layer motivates the treatment of untranslated Urdu material and typographic marking as ideological rather than merely stylistic choices. The fourth layer is evaluation and stance. Identity positions are not merely occupied but appraised; they are affiliated with, distanced from, mourned, or satirized. This layer motivates aspect-level sentiment and stance analysis of identity targets. The AI-corpus pipeline is aligned to these layers: language identification and lexical extraction operationalize indexicality; keyword, collocation, and cluster analysis operationalize distribution; ratio-based and typographic measures

operationalize appropriation; and stance classification operationalizes evaluation. Interpretation proceeds by relating the four layers to one another rather than by reporting them separately.

3. Materials and Methods

3.1. Research Design

This study employed a convergent mixed-methods design with a corpus-analytic core. Quantitative procedures establish distributional facts about the corpus; qualitative procedures interpret those facts through concordance reading and contextual analysis. The design is sequential within each analytical strand (machine analysis followed by human validation and interpretation) but convergent overall, in that lexical, structural, semantic, and evaluative strands are integrated at the modelling stage. The orientation is descriptive and exploratory rather than hypothesis testing: the aim is to characterize a literary formation, not to confirm a prior claim about it.

3.2. Selection of Contemporary Pakistani Fiction

Texts were selected to represent contemporary Pakistani fiction in English as a publishing formation rather than a critical canon. The sampling frame combined three sources: publisher catalogues of Pakistani and South Asian imprints, longlists and shortlists of major literary prizes for which Pakistani writers were eligible, and academic bibliographies of Pakistani Anglophone writing. From this frame, twenty-four titles by twenty authors were selected to balance gender, generation, setting, and mode. Both internationally published and locally published titles were included, as restricting the corpus to internationally circulated books would bias it toward texts written with a foreign readership in view.

3.3. Inclusion and Exclusion Criteria

The inclusion criteria were as follows: (a) prose fiction, whether novel or single-author short-story collection; (b) first publication between 2000 and 2023; (c) original composition in English rather than translation; (d) authorship by a writer of Pakistani origin, whether resident or diasporic; and (e) substantial engagement with Pakistani settings, characters, or communities. The exclusion criteria were as follows: translated fiction, poetry and drama, children's and young-adult fiction, memoir and reportage, anthologies with multiple authors, and texts for which a reliable digital edition could not be obtained. Genre fiction was not excluded on principle, and two crime novels were retained, as genre writing often carries dense vernacular material.

3.4. Corpus Construction

The corpus, designated the Contemporary Pakistani Fiction Corpus (CPFC), was constructed in four stages: acquisition of digital editions, conversion to plain text, structural mark-up, and metadata attachment. Each text file received a header recording the author, title, year of first publication, place of publication, primary setting, narrative person, and author's residence status. Within each file, chapter boundaries, dialogue, and narration were marked so that later analyses could be restricted to particular narrative levels. This structural layer proved essential, since the distribution of multilingual material differed sharply between dialogue and narration.

A reference corpus was required for the keyword analysis. A comparison set of contemporary general fiction in English of approximately 11 million words, matched for period and length distribution and drawn from non-South Asian authors, was used as the reference. Keyness results should therefore be read as 'distinctive relative to contemporary general Anglophone fiction', not as absolute frequency claims.

3.5. Corpus Size and Composition

The final corpus comprises 24 texts, 2,943,168 running words (tokens), and 71,402 distinct word types, with a mean text length of 122,632 words. Dialogue accounts for 34.7% of tokens and narration for 65.3%. Table 1 summarizes the composition by decade, gender of author, residence status, and dominant setting.

Table 1. Composition of the Contemporary Pakistani Fiction Corpus (CPFC)

Category	Sub-category	Texts	Tokens	% of corpus
Period	2000–2009	9	1,043,510	35.5
	2010–2016	8	1,014,238	34.5
	2017–2023	7	885,420	30.1
Author gender	Female	12	1,466,905	49.8

Category	Sub-category	Texts	Tokens	% of corpus
Residence	Male	12	1,476,263	50.2
	Pakistan-resident	13	1,507,912	51.2
	Diasporic	11	1,435,256	48.8
Dominant setting	Karachi	7	902,447	30.7
	Lahore	5	621,338	21.1
	Other Pakistan	6	708,190	24.1
	Transnational	6	711,193	24.2

3.6. Text Digitization and Pre-processing

Where texts were available only in print, pages were scanned at 400 dpi and processed with optical character recognition, followed by manual correction of a 10% sample and automatic detection of residual errors through frequency-anomaly screening. Pre-processing included the removal of front and back matter, normalization of quotation marks and dashes, retention of original diacritics and italics as separate mark-up attributes, sentence segmentation, and tokenization with a tokeniser modified to keep transliterated multiword units such as *shalwar kameez* intact. Crucially, lower-casing was applied only in a parallel copy of the corpus; the primary copy retained case, since capitalization carries information about address terms and proper nouns.

3.7. AI-Based Corpus Analysis Framework

The analytical framework comprises five modules that operate over the same annotated corpus. Module 1 performs surface description: frequency, keyness, dispersion, and n-gram profiling. Module 2 performs syntactic and entity annotation using a transformer-based tagger and named entity recognizer. Module 3 performs multilingual detection: character-level and token-level language identification, dictionary matching against a purpose-built Urdu-Punjabi-Sindhi-Pashto transliteration lexicon, and classification of switch type. Module 4 performs semantic analysis: sentence embedding, clustering, topic modelling, and semantic field assignment. Module 5 performs evaluative analysis: aspect-based sentiment and stance classification over identity targets. Outputs from all modules are written to a unified relational store keyed by text, chapter, sentence, and token identifiers, which permits cross-module queries such as 'stance polarity of sentences containing Urdu kinship terms in dialogue by female characters'.

3.8. Natural Language Processing Procedures

Part-of-speech tagging, lemmatization, and dependency parsing were performed using a transformer-based pipeline. Named entity recognition was adapted to the corpus by fine-tuning on 4,000 manually annotated sentences containing South Asian personal names, place names, and organization names, since off-the-shelf models systematically mis-classified Urdu personal names and Pakistani place names. Coreference resolution was applied to link character mentions across paragraphs, enabling speaker attribution for 82.4 per cent of dialogue turns; the remainder were left unattributed rather than guessed.

3.9. Keyword and Frequency Analysis

Keyword analysis compared CPFC against the reference corpus using log-likelihood for significance and log ratio for effect size, with a minimum frequency threshold of 20 occurrences and a dispersion requirement that a keyword occurs in at least six texts. This dispersion filter is important in literary corpora, where a single long novel can generate spurious keywords through the repetition of a character's name or a local obsession. Function word keyness was examined separately from content word keyness.

3.10. Collocation and Concordance Analysis

Collocates were computed within a five-word window on either side of the node, ranked by logDice, with a minimum co-occurrence frequency of five. For each identity node, the top 30 collocates were inspected in concordance, and collocates were grouped into functional categories by two analysts working independently. Concordance sets exceeding 400 lines were sampled randomly at 200 lines for manual reading. All examples reported in this study were paraphrased or constructed to illustrate patterns; verbatim extended quotations from copyrighted texts are avoided throughout.

3.11. Code-Switching and Multilingual Expression Detection

Detection proceeded in three passes. The first pass applied a character n-gram language identifier trained on Romanised Urdu, Punjabi, Sindhi, Pashto and Arabic material, producing a candidate list of non-English tokens. The second pass matched candidates against a transliteration lexicon of 6,780 entries compiled from existing glossaries, bilingual dictionaries and corpus-driven extraction, which handled orthographic variation such as ammi/ami/ammee. The third pass classified each confirmed switch by structural type (single-word insertion, multiword phrase, full clause, discourse marker, formulaic religious expression) and by narrative location (narration, dialogue, free indirect thought). Precision and recall were estimated against a manually annotated 60,000-word sample.

3.12. Semantic and Thematic Analysis

Sentences were embedded using a multilingual sentence-transformer model to ensure that code-mixed sentences received usable representations. Embeddings were reduced dimensionally and clustered; the clusters were then labelled by extracting class-distinctive terms and by reading a stratified sample of twenty sentences per cluster. In parallel, a semantic-field tagger assigned tokens to categories such as kinship, religion, food, dress, violence, mobility, and law, with an extension dictionary added for South Asian items absent from the default resource.

3.13. Sentiment and Stance Analysis

Aspect-based sentiment classification was applied to sentences containing identity targets, with targets defined as mentions of nations, cities, ethnolinguistic groups, religious categories, gender categories, and social classes. A three-way polarity scheme (positive, negative, neutral) was combined with a stance scheme distinguishing affiliation, distancing, ambivalence, and critique. The stance classifier was fine-tuned on 2,500 manually annotated sentences from the corpus. Because sentiment models are known to be culturally biased, all aggregate polarity claims were checked against manual coding of a 500-sentence random sample.

3.14. Linguistic Identity Feature Extraction

Twenty-two features were extracted per text and normalized per 10,000 words. These included: borrowing density; ratio of untranslated to glossed borrowings; proportion of italicized to non-italicised borrowings; kinship-term density; religious-formula density; address-term inventory size; first-person plural pronoun density; ratio of 'we' to 'they'; density of place names; ratio of Pakistani to non-Pakistani place names; density of language-referring metalanguage such as 'accent', "Urdu", "English", "mother tongue"; dialogue proportion; and stance ratios for each identity target class. These features formed the input to the comparative analyses in Sections 4.12 to 4.15.

3.15. AI-Assisted Topic Modelling

Topic modelling was performed on paragraph-level units using an embedding-based model with class-based term weighting and cross-checked with a probabilistic model at comparable topic numbers. Solutions from five to twenty-five topics were generated; coherence scores and human interpretability ratings converged on a nine-topic solution. Topics were labelled by three raters independently, with disagreements resolved in discussion; inter-rater agreement on topic labels before discussion was 0.79.

3.16. Validation of AI-Generated Results

Four validation procedures were applied. First, module-level accuracy testing against manually annotated samples, reported in Table 2. Second, triangulation between methods, so that a theme was accepted only if supported by at least two of keyword, collocation, cluster, and topic evidence. Third, concordance-level verification, in which every quantitative claim reported in the Results was checked against at least twenty concordance lines. Fourth, negative-case analysis, in which texts departing from the general pattern were examined individually rather than treated as noise.

Table 2. Validation results for AI modules against manually annotated samples

Module	Sample size	Precision	Recall	F1 / Agreement
POS tagging (adapted)	8,000 tokens	0.968	0.964	0.966
Named entity recognition	6,500 tokens	0.913	0.887	0.900
Code-switch detection	60,000 tokens	0.941	0.902	0.921
Switch-type classification	1,800 switches	0.886	0.871	0.878

Module	Sample size	Precision	Recall	F1 / Agreement
Semantic field assignment	5,000 tokens	0.874	0.851	0.862
Topic label agreement	9 topics × 3 raters	-	-	0.79 (pre-discussion)
Stance classification	2,500 sentences	0.842	0.828	0.835

3.17. Qualitative Interpretation of Corpus Findings

The quantitative output was interpreted through a structured reading protocol. For each pattern, the analysts recorded: the node or feature; its distribution across texts, narrative levels, and character types; representative contexts; competing interpretations; and the textual evidence that discriminated between them. The interpretations were then related to the four theoretical layers set out in Section 2.11. Where the corpus evidence and critical literature diverged, the divergence was reported rather than reconciled.

3.18. Reliability and Validity

Reliability was addressed through documented pre-processing, fixed random seeds for stochastic procedures, double annotation of validation samples with reported agreement coefficients, and preservation of the full analytical pipeline as executable scripts. Validity was addressed through construct definition prior to feature extraction, dispersion filtering to prevent single-text artifacts, triangulation across methods, and explicit acknowledgement that keyness is relative to the chosen reference corpus. The principal residual threat to validity is representativeness: twenty-four texts cannot exhaust a literary field, and the exclusion of Urdu-language fiction means that the conclusions apply to Anglophone writing only.

3.19. Ethical and Copyright Considerations

All the analyzed texts were legitimately acquired. The corpus was used for non-consumptive computational analysis, and no substantial portion of any copyrighted text was reproduced here; illustrative examples were paraphrased, constructed, or limited to individual lexical items and short collocational patterns. The corpus itself is not redistributed; instead, derived data, such as frequency lists, collocation tables, and annotation schemes, are made available for verification. No human participants were involved; therefore, ethics approval for human subject's research was not required. However, the study observes the principle that authors' work should be represented accurately and that findings about identity should not be used to essentialize the communities depicted.

4. Results

4.1. Descriptive Characteristics of the Fiction Corpus

The corpus contains 2,943,168 tokens and 71,402 types, giving a standardized type-token ratio of 46.2 per 1,000 words, slightly higher than the reference corpus value of 44.1, consistent with the higher lexical variety introduced by multilingual material and place-specific vocabulary. The mean sentence length is 17.9 words, close to the reference value of 17.4. Dialogue occupies 34.7% of the corpus, but contains 71.3% of all detected non-English tokens, an early indication that multilingual identity work is concentrated in represented speech. Non-English material accounts for 1.62% of running tokens overall, or 162 tokens per 10,000 words. This proportion is small in aggregate but highly clustered: the interquartile range across texts runs from 71 to 248 per 10,000 words, and the most vernacular text records 419 per 10,000 words against 23 per 10,000 in the least.

4.2. High-Frequency Linguistic Patterns

Beyond expected function words, the most frequent content lemmas in the corpus are man, house, mother, father, city, night, water, God, name, hand, country, road, wall, and door. The prominence of house, wall, door, and road reflects a persistent concern with enclosure, threshold, and passage that recurs in later topic analysis. The lemma God is markedly frequent relative to the reference corpus, and concordance inspection shows that a substantial proportion of its occurrences are formulaic rather than doctrinal. Four-gram analysis reveals a set of habitual frames with identity function: 'in the name of', 'the rest of the family', 'as if she were', 'the sound of the', 'in those days we', 'a woman like that' and 'people like us'. The last two are of particular interest, since both are categorization devices: they invoke social types without naming them, leaving the reader to supply the relevant classification. Their combined frequency is 3.1 per 100,000 words, against 0.9 in the reference corpus.

4.3. Keywords Associated with Pakistani Identity

Keyword analysis returned 1,412 items above the threshold. After removing character and place names, the strongest content keyword cluster was found in six fields, as presented in Table 3.

Table 3. Selected keywords by semantic field (log-likelihood, log ratio against the reference corpus)

Semantic field	Representative keywords	Texts (of 24)	LL	Log ratio
Kinship address	and ammi, abba, beta, baji, khala, chacha, sahib, bibi	22	8,412.6	6.94
Religion morality	and inshallah, mashallah, namaz, maulvi, kafir, azaan, haram	21	6,978.1	6.51
Food and domestic life	biryani, chai, paratha, achar, dhaba, charpai, mohalla	20	5,631.4	6.22
Dress and body	dupatta, kameez, shalwar, kurta, chappal, burqa	19	5,104.9	6.05
Place and polity	Karachi, Lahore, Punjab, cantonment, katchi, partition	24	9,220.3	5.78
Conflict and order	curfew, checkpoint, blast, thana, martial, mullah, mohajir	18	4,286.7	5.43

Two observations follow. First, the keyword profile is overwhelmingly nominal: identity in this corpus is lexicalized through things, persons, and places rather than through distinctive verbs. Second, the fields are unevenly dispersed. Place-and-polity keywords occur in every text, whereas conflict-and-order keywords are concentrated in 18 texts and are strongly associated with post-2010 publication.

4.4. Collocational Patterns of Identity-Related Terms

Six identity nodes were analyzed in detail: home, country, people, name, blood, and language. The node home shows a bifurcated collocational profile. On one side are possessive and deictic collocates (my, our, back, went, came, left) that treat home as a location; on the other are modifiers and predicates of loss and doubt (real, ever, called, supposed, never, again, was not) that treat home as a claim under question. The second group accounts for 38% of significant collocates, a proportion without parallel in the reference corpus, where home collocates predominantly with movement verbs. The node country collocates with left, whole, this, own, back, serve, loved, ruined, betrayed, and love. The co-presence of loved and betrayed within the same collocational field, frequently in the same paragraph, is the clearest single indicator of the ambivalence that characterizes national identity language in the corpus.

The node people is strongly associated with categorizing modifiers: our people, these people, those people, poor people, rich people, city people, and village people. In 61% of its occurrences, people is preceded by a determiner or possessive that performs group boundary work rather than a mere reference. The node name attracts change, changed, real, family, father's, mean, means, wrote and pronounce. Concordance evidence shows a recurrent narrative situation in which a character's name is mispronounced, shortened, or replaced in a foreign or official setting, and the passage explicitly comments on the loss. Thus, name is a metalinguistic site of identity negotiation. The node blood divides between literal violence collocates (pool, dried, stains, on his) and metaphorical descent collocates (own, in our, thicker, of the family). The metaphorical share is 29 percent, notably higher than in the reference corpus, and rises to 41 percent in texts with Partition or inheritance plots. The node language collocates with mother, own, secret, common, learn, forget, lost, English, and Urdu. The pairing of forget and lost with mother tongue is frequent enough to constitute a formula of diasporic self-description.

4.5. Code-Switching and Multilingual Identity Markers

A total of 47,662 non-English tokens were detected, forming 31,884 switch events. Table 4 presents the distribution by structural type and narrative location.

Table 4. Distribution of switch events by structural type and narrative location

Switch type	Events	% of switches	In dialogue (%)	In narration (%)
Single-word noun insertion	18,742	58.8	62.4	37.6
Multiword phrase insertion	5,118	16.1	70.1	29.9
Address / kinship term	3,905	12.2	91.7	8.3
Religious formula	2,441	7.7	88.2	11.8
Discourse marker (na, yaar, acha, arre)	1,268	4.0	97.6	2.4
Full clause switch	410	1.3	96.1	3.9

Single-word noun insertion dominates, and it is the only type that occurs substantially in narration: the narratorial voice borrows the names of things but rarely the rhythms of speech. Discourse markers and full clauses are almost entirely confined to the dialogue. This asymmetry is theoretically important. This indicates a division of labor in which narration performs referential localization and dialogue performs interactional authentication. Typographic marking declined across the period. In 2000–2009 texts, 68.4 per cent of borrowings are italicized; in 2017–2023 texts, only 24.1 per cent are. In-text glossing shows the same trajectory, falling from 31% to 9% of first occurrences. Later texts more often let unfamiliar items stand unexplained, addressing a reader assumed to be either competent or willing to remain outside.

4.6. Urdu and Regional-Language Lexical Borrowings

Of confirmed non-English types, 74.2 per cent are attributable to Urdu, 9.6 per cent to Punjabi, 5.1 per cent to Arabic used as religious register, 4.3 per cent to Sindhi, 3.4 per cent to Pashto, 2.2 per cent to Persian and 1.2 per cent to other sources. Regional-language items are strongly text-specific: Pashto borrowings occur almost exclusively in texts set in Khyber Pakhtunkhwa or the tribal districts, and Sindhi borrowings cluster in two Sindh-set novels. In contrast, Urdu borrowings are dispersed across all twenty-four texts. Semantically, borrowings concentrate in kinship (21.7 per cent of borrowed tokens), food and drink (16.4), religion (15.8), dress (11.2), address and honorifics (10.9), architecture and domestic space (8.6), abuse and exclamation (6.3), law and officialdom (4.8), and other (4.3). The near-absence of borrowing in the fields of technology, finance, and abstract argument is as informative as its presence elsewhere: the vernacular is assigned to intimacy, embodiment, and belief, while English retains the domains of system and abstraction. This distribution reproduces at the level of style the diglossia hierarchy described in the language-policy literature.

4.7. Cultural and Religious Lexical Expressions

Religious expressions occurred at 41.2 per 10,000 words. Manual coding of a 600-instance sample found that 57 percent are phatic or interactional (marking hope, gratitude, warning, or closure of a topic), 24 percent are characterizing (assigning a speaker to a social or ideological position), 13 percent are ironic or critical, and only 6 percent are doctrinal. This finding qualifies a common critical assumption that religious lexis in this fiction primarily signals religiosity; distributionally, it more often signals sociability. Density is not evenly spread across characters. In texts where speaker attribution was recoverable, religious formulae are 2.4 times more frequent in the speech of older characters than younger ones, and 1.9 times more frequent in the speech of characters marked as rural or lower-middle-class than in that of elite urban characters. Where elite characters use them, concordance reading shows a high proportion of ironic or performative uses.

4.8. Pronouns and Collective Identity Construction

First-person plural forms (we, us, our, ourselves) occur at 78.3 per 10,000 words, against 61.5 in the reference corpus. Third-person plural forms show a smaller difference. However, the most informative measure is the ratio of exclusive to inclusive we, determined by manual coding of a 700-instance sample: 64% of instances are exclusive, defining a group that does not include the addressee or implied reader. Exclusive we is most frequent in three environments: comparisons with the West, discussions of family obligation, and passages of collective memory concerning Partition, war, or dictatorship. In each case, the pronoun performs boundary work without specifying the basis of the boundary, allowing national, religious, class, and familial belongings to be invoked simultaneously. The pattern 'people like us' noted in Section 4.2 is a lexicalized extension of the same mechanism, and its concordance shows that it is almost always used to mark class rather than nation.

4.9. Stance and Evaluation Patterns

The aspect-level classification of 42,318 sentences containing identity targets produced the distribution reported in Table 5.

Table 5. Stance distribution across identity targets (percentages of classified sentences)

Identity target	Sentences	Affiliation	Ambivalence	Distancing	Critique
Nation (Pakistan)	9,614	21.8	38.6	12.4	27.2
City (Karachi, Lahore, others)	8,220	34.1	31.7	13.9	20.3
Family and kin group	10,455	42.6	29.8	14.5	13.1
Religious community	6,112	24.9	27.4	18.2	29.5
Gender categories	4,908	18.3	22.1	17.9	41.7
Social class	3,009	11.4	19.6	26.8	42.2

Three findings stand out. First, ambivalence rather than affiliation is the modal stance toward the nation, confirming at scale what individual readings have long asserted. Second, the city attracts markedly more affiliation than the nation: urban belonging is more securely held than national belonging. Third, gender and class attract the highest proportions of critique, indicating that these are the categories the corpus most consistently treats as objects of contestation rather than grounds of identification.

4.10. AI-Detected Semantic Themes

Embedding-based clustering of 128,400 paragraph units produced fourteen initial clusters, of which eleven were interpretable and nine were retained after merging near-duplicates. The retained clusters concern: domestic interiors and family hierarchy; religious practice and moral judgement; migration, airports, and documents; political violence and security infrastructure; class relations between employers and servants; romantic and marital negotiation; childhood, school, and language learning; landscape, weather, and rural labor; and media, rumor, and public speech.

4.11. Topic-Modelling Results

The nine-topic solution is summarized in Table 6, with topic prevalence expressed as a percentage of the paragraph units.

Table 6. Nine-topic solution with distinctive terms and prevalence

Topic	Label	Distinctive terms	Prevalence (%)
T1	Family and domestic order	mother, house, kitchen, ammi, marriage, room	17.4
T2	Migration and belonging	visa, airport, back, home, foreign, letters	13.9
T3	Religion and moral order	prayer, namaz, sin, maulvi, mosque, forgiveness	12.1
T4	Urban violence and security	police, blast, curfew, thana, target, street	11.6
T5	Class and service relations	servant, master, wages, village, landlord, kothi	10.8
T6	Gender, marriage and honour	girl, honour, husband, veil, shame, mother-in-law	10.2
T7	Language, school and English	English, Urdu, accent, convent, teacher, tuition	8.7
T8	History, war and the state	partition, army, general, martial, refugee, border	8.4
T9	Landscape, weather and labour	canal, dust, monsoon, wheat, heat, mango	6.9

Topics T2, T3, T6, and T7 are the most productive of identity language, measured by the density of identity keywords and identity-target sentences within them. T7 is the smallest of these but the most concentrated: paragraphs in this topic contain metalinguistic vocabulary (accent, mother tongue, fluent, broken) at eleven times the corpus average. Language itself, in other words, is an explicit thematic object and not merely the medium of fiction.

4.12. Identity Patterns across Authors and Novels

Cluster analysis of the 22 normalized features grouped the texts into three profiles. Profile A (nine texts) is characterized by high borrowing density, low italicization, high dialogue proportion, and high exclusive-we ratio; these are predominantly Pakistan-resident authors writing in local settings. Profile B (eight texts) shows moderate borrowing, high glossing, high place-name diversity including non-Pakistani locations, and elevated migration-topic prevalence; these are predominantly diasporic-authored transnational narratives. Profile C (seven texts) shows low borrowing, low dialogue proportion, and high abstraction, with identity work carried by pronouns and metalanguage rather than by lexis; these are formally experimental or second-person and essayistic narratives. Residence status is the strongest single predictor of Profile A membership, but it is not deterministic: two diasporic-authored texts fall into Profile A, and both are set entirely within Pakistan, suggesting that setting exerts more influence on vernacular density than does the author's location.

4.13. Gender-Related Linguistic Identity Patterns

Where character gender could be attributed, female characters' dialogue contained 1.6 times more kinship terms and 1.4 times more religious formulae than male characters' dialogue, while male characters' dialogue contained 2.1 times more abusive and exclamatory borrowings. Female characters were also more frequently the targets of sentences classified as critiques of gender norms (68.3% of such sentences), but the critique was generally directed at the norm rather than at the character. Texts by female authors show a higher density of domestic-interior vocabulary and a higher prevalence of Topic T6, while texts by male authors show higher prevalence of T4 and T8. The difference is one of degree rather than kind; the interquartile ranges overlap substantially, and two of the three texts with the highest T4 prevalence are by women.

4.14. Social-Class and Cultural Identity Patterns

Class is realized linguistically through a consistent contrast. Elite characters' speech is rendered in near-standard English with occasional performative vernacular items, often marked as knowing; subaltern characters' speech is rendered either in translated English with syntactic transfer or in dense vernacular. Address terms carry much of the load: sahib, begum, bibi, and baba appear at 9.4 per 10,000 words in the corpus, and 79 per cent of their occurrences are in asymmetrical exchanges between employer and employee or landlord and tenant. The stance data reinforce this observation. Social class attracted the highest critique proportion of any target (42.2 percent) and the highest distancing proportion (26.8 percent). Concordance reading suggests that critique is generally directed upward at elite complacency, while distancing is more often produced by narrators positioning themselves against the milieu they describe.

4.15. Urban, Diasporic, and National Identity Patterns

Karachi and Lahore have distinguishable linguistic signatures. Karachi-set passages are keyed for mohajir, target killing, katchi abadi, sea, traffic, and Urdu-language markers; Lahore-set passages are keyed for Punjabi lexis, walled city, canal, mangoes, and forms of address associated with older elite networks. Both are affiliated with the nation more warmly than the nation is. Diasporic passages show a distinct profile: elevated metalinguistic vocabulary, elevated frequency of name and accent, high density of documents and institutions (visa, passport, immigration, form), and a marked increase in the ratio of they to we. National identity language, by contrast, is concentrated in retrospective and collective passages and is dominated by the ambivalence pattern documented in Section 4.9. Across the corpus, the nation is spoken of most often in the past tense and the city most often in the present.

4.16. Integrated AI-Corpus Model of Linguistic Identity

The findings can be integrated into a four-tier model: Tier 1, the lexical tier, comprises borrowings, address terms, religious formulae and place names. It performs referential localization and is measurable through keyness and the borrowing density. Tier 2, the structural tier, comprises switch types, typographic marking, glossing, and syntactic transfer. It performs authentication and reader positioning and is measurable through switch classification and marking ratios. Tier 3, the discursive tier, comprises pronominal boundary work, categorization devices, and metalinguistic commentary; it performs group construction and is measurable through pronoun ratios and metalanguage density. Tier 4, the evaluative tier, comprises stance toward identity targets; it appraises the positions constructed in the lower tiers and is measurable through aspect-based stance classification. The tiers are not independent. High Tier 1 density with low Tier 2 marking produces the naturalized vernacular of Profile A texts; high Tier 1 density with high Tier 2 marking produces the explanatory mode of earlier diasporic writing; and low Tier 1 density with high Tier 3 activity produces the abstract identity work of Profile C. Tier 4 modulates all of these, so that the same lexical repertoire can support affiliation in one text and satire in another. The model's claim is that a text's identity signature is the joint configuration of the four tiers, and that this configuration can be measured.

5. Discussion

5.1. Major Linguistic Identity Patterns in Pakistani Fiction

The corpus evidence supports a compact characterization of how contemporary Pakistani fiction in English constructs identity. Identity is lexicalized nominally, concentrated in the domains of kinship, faith, food, dress, and place; it is performed interactionally in dialogue, where switching and address terms carry most of the weight; it is bounded pronominally through an unusually high proportion of exclusive first-person plural forms; and it is appraised ambivalently, with the nation attracting more ambivalence than affiliation while the city and the family attract more affiliation than either. The most consequential single finding is the division of labor between narration and dialogue. Narration borrows nouns, while dialogue borrows rhythms. This division suggests that these texts distinguish two kinds of authenticity: the authenticity of the described world, which can be signalled by naming things correctly, and the authenticity of the represented person, which requires the reproduction of pragmatic habits. Identity, in the corpus, resides more securely in the second than in the first.

5.2. Multilingualism as a Marker of Pakistani Identity

The distribution of borrowings by semantic field is important for diagnosis. Vernacular material is dense where the body, household, and belief are concerned and almost absent where systems, abstractions, and modern institutions are concerned. This is not a neutral, aesthetic preference. It reproduces the domain allocation that language policy has produced in society, where English holds administration, higher education, and the formal economy, while Urdu and regional languages hold intimacy, ritual, and the street. Thus, Anglophone Pakistani fiction does not simply describe a diglossic society; its own lexical economy is organized diglossically. This has critical implications. Celebratory accounts of hybridity sometimes treat multilingual insertion as inherently subversive of colonial linguistic hierarchy. The corpus suggests a more qualified conclusion: insertion localizes and enriches, but it operates within a division of labor that leaves the prestige domains of English untouched.

Subversion, where it occurs, is more evident in the decline of italicization and glossing than in the presence of borrowings as such, because refusing to mark or explain is a claim about who the text's reader is.

5.3. Code-Switching and Hybrid Identity Construction

Structurally, the switching in the corpus is overwhelmingly of the embedded-content type: English supplies the morphosyntactic frame, and Urdu supplies nouns, address terms, and formulae. Full clause switches are rare, and where they occur, they are usually accompanied by an English rendering nearby, indicating that they are dramatic rather than communicative in function. Pragmatically, switching performs at least five identifiable functions in the corpus: marking intimacy, where a shift to kinship or endearment terms signals proximity; marking authority, where religious or legal formulae are used to close an argument; marking exclusion, where a switch signals that a third party is not addressed; marking irony, where a formula is used by a character who does not endorse it; and marking untranslatability, where the text asserts that the concept has no English equivalent. The latter is the most theoretically loaded, as it converts a lexical gap into an identity claim. Hybridity, then, is not a uniform condition in these texts but a set of graded techniques. The corpus permits the gradation to be measured, which is precisely what illustrative criticism cannot do.

5.4. Local Lexical Choices and Cultural Representation

The typographic findings deserve further emphasis. The fall in italicization from roughly two-thirds to roughly one-quarter of borrowings across two decades marks a change in the implied reader's expectations. Italics frame a word as foreign to the text's default language; its removal asserts that the word is at home. Similarly, the decline in glossing indicates a reduced willingness to translate for an external audience. Both changes are consistent with the increasing confidence of a literary formation that no longer regards the metropolitan reader as a necessary destination. This should not be over-read. Publishers, house styles, and editorial intervention influence typography, and the study cannot separate authorial from editorial decisions. What the corpus establishes is that the change happened, not who caused it.

5.5. Linguistic Construction of National and Collective Identity

Pronominal evidence indicates that collective identity in the corpus is built through unspecified boundary work. The high proportion of exclusive *we* and the productivity of categorization devices such as *people like us* allow texts to invoke belonging without committing to its basis. This vagueness is functional; it permits national, religious, class, and familial affiliations to be layered and allows a single utterance to be read differently by readers positioned differently. The stance results add a further dimension. Ambivalence toward the nation is not merely a thematic mood but a measurable evaluative pattern, present across authors of both genders and both residence statuses, and stable across the three periods. Where affiliation is expressed, it attaches more readily to the city, the neighborhood, and the family than to the state. The corpus thus supports an account of Pakistani literary identity as strongly localist and weakly statist.

5.6. Negotiation of Local and Global Identities

Diasporic passages exhibit a distinct linguistic economy: heavy metalinguistic vocabulary, documents and institutional nouns, and a rise in the ratio of *they* to *we*. Identity here is negotiated at points of institutional friction, such as the airport, visa office, and classroom, and the recurring motif of the mispronounced or shortened name is the compressed form of this negotiation. Importantly, the corpus does not support a simple opposition between rooted local texts and deracinated global texts. Two diasporic-authored texts pattern with the most vernacular local texts, and several Pakistan-resident texts show high metalinguistic density. Setting and narrative situation predict linguistic profile better than author biography does, which argues against biographical readings of linguistic hybridity.

5.7. Comparison with Previous Studies

The findings converge with descriptive work on Pakistani English in identifying kinship, religion, food, and dress as the principal borrowing domains, and they corroborate critical readings that emphasize ambivalence toward the nation. They diverge from parts of the critical literature in three respects. First, religious lexis emerges as predominantly phatic rather than doctrinal, which complicates readings that treat its frequency as evidence of religious preoccupation. Second, code-switching is far more concentrated in dialogue than general accounts of textual hybridity imply, so claims about hybridized narration require qualification. Third, the corpus does not support a strong gender dichotomy in thematic concern; the differences observed are gradient and overlapping, and the strongest single predictor of linguistic profile is setting.

5.8. Contribution of AI-Based Corpus Analysis

The AI components contributed in ways that classical corpus tools could not. Language identification combined with a transliteration lexicon made systematic switch detection feasible across orthographic variations that would defeat a fixed word list. Embedding-based clustering grouped passages that shared meaning without sharing vocabulary, which was essential for

identifying the metalinguistic theme. Aspect-based stance classification allowed evaluation to be attached to specific identity targets rather than to whole sentences, producing the differentiated picture in Table 5. Fine-tuned entity recognition made speaker attribution possible at a scale that permitted character-level comparison. The limits were equally instructive. Off-the-shelf sentiment models misclassified formulaic religious expressions as strongly positive and rendered ironic usage invisible; only manual sampling revealed the error. Topic solutions were sensitive to unit size, and paragraph-level units proved necessary because sentence-level units fragmented scenes. Named entity recognition required domain adaptation before it was usable. None of these problems is exotic; all of them silently distort results in a pipeline without validation. The methodological lesson is that AI methods extend reach but do not reduce the analyst's obligation to read.

5.9. Theoretical Implications

For sociolinguistic theory, the study offers large-scale textual support for a positioning account of identity and adds a specification: in literary discourse, positioning is distributed across tiers that can be independently measured, and a text's identity signature is their joint configuration. For corpus stylistics, the study demonstrates that semantic-level annotation makes it possible to move from keyword lists to functional claims, provided that validation is built into the design. For postcolonial theory, the study suggests that appropriation is better measured by marking and glossing behavior than by borrowing density, since the ideological claim resides in whether the text explains itself.

5.10. Implications for Pakistani Literary and Linguistic Studies

Three practical implications follow. For literary studies, corpus evidence provides a way to test period claims about the field rather than generalizing from three or four canonical novels, and it identifies underexamined features such as address-term asymmetry and metalinguistic density as productive objects of interpretation. For linguistics, the corpus documents literary Pakistani English as a describable register with quantified borrowing profiles, useful for comparison with journalistic and conversational data. For pedagogy, the annotated corpus supports data-driven teaching in stylistics and translation studies, allowing students to test hypotheses about hybridity against evidence.

6. Conclusion

This study analyzed a 2.94-million-word corpus of 24 works of contemporary Pakistani fiction in English using an AI-assisted corpus-analytic framework. Six principal findings emerged. First, identity is lexicalized nominally and concentrated in kinship, religion, food, dress, and place. Second, multilingual material occupies 1.62% of tokens overall but 71.3% of it occurs in dialogue, establishing a division of labor in which narration localizes reference while dialogue authenticates persons. Third, typographic marking and glossing of borrowings declined sharply across the period, indicating a shift in the implied reader. Fourth, core identity nodes show ambivalent collocational profiles, with country associating simultaneously with love and betrayal and home associating heavily with doubt. Fifth, exclusive first-person plural forms and categorization devices perform boundary work whose basis is left unspecified. Sixth, stance analysis shows ambivalence as the modal position toward the nation, affiliation toward city and family, and critique toward gender and class. Empirically, the study delivers the first systematically documented corpus-based description of linguistic identity across contemporary Pakistani Anglophone fiction as a formation, together with quantified profiles of borrowing, switching, pronominal, and stance behavior. Theoretically, it proposes a four-tier integrated model in which lexical, structural, discursive, and evaluative layers jointly constitute a text's identity signature, and it shows how each tier can be operationalized.

6.1. Methodological Contributions

This study specifies a reusable pipeline for AI-based literary corpus analysis in multilingual postcolonial contexts, including a transliteration lexicon for switch detection across orthographic variation, a switch-type and switch-location classification scheme, an identity-target stance annotation scheme, and a four-part validation architecture combining module accuracy testing, cross-method triangulation, concordance verification, and negative-case analysis. It also documents specific failure modes of standard NLP tools on this material, which future studies can anticipate rather than rediscover.

7. Limitations

Five limitations should be borne in mind. The corpus contains 24 texts and cannot represent the full field; the results are indicative of tendencies rather than exhaustive. The reference corpus was general contemporary fiction, so keyness results are relative to that comparison and would change against a South Asian reference. Speaker attribution succeeded for 82.4% of dialogue turns; therefore, character-level findings exclude a non-random remainder, likely biased toward crowded scenes. Sentiment and stance classification, despite fine-tuning and manual checking, remains the least reliable module, particularly for irony. Finally, the study covers only Anglophone fiction; Urdu-language and regional-language fiction would almost certainly yield a different, and in some respects opposite, picture of the same linguistic ecology.

8. Recommendations for Future Research

Six directions were recommended. First, the corpus should be expanded to fifty or more texts with a balanced representation of locally published fiction. Second, a parallel Urdu-language fiction corpus should be constructed to permit a cross-linguistic comparison of identity construction within the same literary culture. Third, diachronic extension backwards to pre-2000 writing was conducted to test whether the decline in glossing is a longer trend. Fourth, the development of domain-adapted models for South Asian literary English, including a sentiment model trained on ironic religious usage. Fifth, reader-response work tests whether the reduction in glossing produces the interpretive effects hypothesized here. Sixth, the extension of the four-tier model to other Anglophone postcolonial literatures would establish whether the tier configurations identified are specific to the Pakistani case or general to the situation of writing in an inherited language.

References

- Anthony, L. (2022). AntConc (Version 4.1) [Computer software]. Waseda University.
- Ashcroft, B., Griffiths, G., & Tiffin, H. (2002). *The empire writes back: Theory and practice in post-colonial literatures* (2nd ed.). Routledge.
- Baker, P. (2006). *Using corpora in discourse analysis*. Continuum.
- Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh University Press.
- Baumgardner, R. J. (Ed.). (1993). *The English language in Pakistan*. Oxford University Press.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Blommaert, J. (2010). *The sociolinguistics of globalization*. Cambridge University Press.
- Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, 7(4–5), 585–614.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). Association for Computational Linguistics.
- Fairclough, N. (1992). *Discourse and social change*. Polity Press.
- Firth, J. R. (1957). *Papers in linguistics 1934–1951*. Oxford University Press.
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794.
- Gumperz, J. J. (1982). *Discourse strategies*. Cambridge University Press.
- Hardie, A. (2012). CQPweb: Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), 380–409.
- Honnibal, M., & Montani, I. (2020). spaCy: Industrial-strength natural language processing in Python [Computer software].
- Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of ICWSM 2014*.
- Kachru, B. B. (1986). *The alchemy of English: The spread, functions and models of non-native Englishes*. Pergamon Press.
- Mahboob, A. (2009). English as an Islamic language: A case study of Pakistani English. *World Englishes*, 28(2), 175–189.
- Mahlberg, M. (2013). *Corpus stylistics and Dickens's fiction*. Routledge.
- Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.
- Myers-Scotton, C. (1993). *Duelling languages: Grammatical structure in codeswitching*. Clarendon Press.
- Norton, B. (2013). *Identity and language learning: Extending the conversation* (2nd ed.). Multilingual Matters.
- Rahman, T. (1990). *Pakistani English: The linguistic description of a non-native variety of English*. National Institute of Pakistan Studies.
- Rahman, T. (1996). *Language and politics in Pakistan*. Oxford University Press.
- Rahman, T. (2002). *Language, ideology and power: Language learning among the Muslims of Pakistan and North India*. Oxford University Press.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 3982–3992).
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Blackwell.
- Wetherell, M. (2010). The field of identity studies. In M. Wetherell & C. T. Mohanty (Eds.), *The SAGE handbook of identities* (pp. 3–26). SAGE.