Critical Review of Social Sciences and Humanities
Volume 6, Issue 1, 2026

Voices Across Scripts: A Protocol for a Multilingual Urdu–Punjabi Suicide Discourse Corpus and the Cultural Calibration of Psycho-Forensic Linguistic Detection

Afshan Ishfaq[1]
Nida Sultan[2]

ABSTRACT

Psycho-Forensic Linguistic Surveillance Initiative (PFLSI) established, respectively, through series papers 1, 2 and 3 of the a small culturally rich corpus of Pakistani suicide notes (N=140) and a large-scale, statistically rigorous English-language Suicide Discourse Corpus (452,000 tokens) (Ishfaq & Sultan, 2026a). Series Paper 3 unified these findings theoretically and specified the seven-layer PFLS pipeline architecture (Ishfaq et al., 2026), identifying its Layer L3 Cultural Calibration module as dependent on lexical resources that do not yet exist at corpus scale: a validated, quantitatively analyzed body of Urdu- and Punjabi-medium suicidal discourse. This paper addresses that dependency directly. It is a corpus-construction and protocol paper, not a results paper: it specifies the design of a Multilingual Suicide Discourse Corpus (M-SDC) comprising native-script Urdu and Punjabi (Shanmukhi) text together with Roman-Urdu code-switched digital discourse, sourced through ethically governed partnerships with Pakistani crisis helplines, moderated mental-health forums, and re-digitized original-language transcription of the Series Paper 1 note corpus. We specify target corpus scale (a proposed 500,000-token Urdu/Punjabi Suicide Discourse Corpus and a matched 500,000-token control corpus), a code-switch annotation scheme adapted from the CALCS shared-task tradition, an izzat-register lexicon derived inductively from Series Paper 1's qualitative findings, and an inter-annotator reliability protocol targeting Cohen's kappa above 0.75 for semantic-domain tags. Because data collection has not yet been completed at the time of writing, we present anticipated findings as a set of pre-registered, directional hypotheses rather than as reported results, and we provide placeholder result tables whose structure mirrors that of Series Paper 2 so that outcomes will be directly comparable across languages once collection concludes. We contend that the absence of such a corpus limits the empirical validation of the PFLS cultural calibration layer and constrains automated detection in linguistically underrepresented populations.

Key Words: *multilingual corpus linguistics; Urdu NLP; Punjabi NLP; code-switching; suicide risk detection; psycho-forensic linguistics; cultural calibration; PFLSI*

[1] Head of Academics and Student Affairs, Institute of Law, Lahore
[2] Lecturer in English, Namal University, Mianwali, Pakistan

Critical Review of Social Sciences and Humanities Volume 6, Issue 1, 2026

1. INTRODUCTION

1.1 From Theoretical Architecture to Empirical Dependency

Series Paper 3 of the PFLSI programme closed with a specific and unresolved obligation. Having unified seven psychological and linguistic frameworks and specified the seven-layer architecture of the Psycho-Forensic Linguistic Surveillance (PFLS) pipeline, it identified what it termed the izzat problem: the existence of culturally embedded registers of suicidal crisis honor-shame vocabulary, domestic entrapment metaphors, relational guilt organized around family obligation that are documented qualitatively in Series Paper 1's Pakistani note corpus but have never been subjected to the same quantitative, corpus-scale scrutiny that Series Paper 2 applied to English-language suicidal discourse (Ishfaq & Sultan, 2026a). The pipeline's Layer L3 Cultural Calibration module was specified architecturally an Urdu/Punjabi sub-model, a metaphor detection component, a code-switching normalizer but none of these components can be built, trained, or validated without the corpus that would give them empirical content. This paper proposes the methodological framework required for constructing that corpus.

1.2 The Monolingual Ceiling

A close reading of the PFLSI programmer's first two empirical papers reveals an asymmetry that Series Paper 3 acknowledged but did not resolve. Series Paper 2's methodological rigor log-likelihood Keynes analysis, USAS semantic domain profiling, Mutual Information collocational analysis, Cohen's d effect sizes against a matched control corpus were applied exclusively to a 452,000-token English-language corpus drawn from four Western-registered sources (Ishfaq & Sultan, 2026a). Series Paper 1's culturally distinctive findings, by contrast, rest on thematic coding and qualitative frequency counting of a far smaller corpus (N=140 notes), processed through Sketch Engine in a manner that the original paper does not specify as operating on native-script Urdu or Punjabi text rather than on translated or transliterated versions. The result is what these paper terms the monolingual ceiling: the programmer's most statistically authoritative evidence exists only for English, while its most culturally specific evidence exists only at small scale and without the corpus-linguistic apparatus that would let it travel into a computational pipeline on equal footing. Until this asymmetry is corrected, any claim that the PFLS pipeline is 'culturally calibrated' describes an aspiration encoded in an architecture diagram, not a property demonstrated in data.

1.3 Statement of the Problem

Three specific and interlocking gaps motivate the present paper. First, there is no publicly documented, quantitatively analyzed corpus of native-script Urdu or Punjabi suicidal discourse comparable in scale or methodological rigor to Series Paper 2's English-language Suicide Discourse Corpus (Ishfaq & Sultan, 2026a). Without such a corpus, claims about the cross-linguistic robustness of the psychache, hopelessness, and interpersonal-theory markers identified in English cannot be tested, and the izzat-register findings of Series Paper 1 cannot be assigned effect sizes, Keynes statistics, or semantic-domain proportions (Ishfaq & Sultan, 2026a).

Second, the computational infrastructure that Series Paper 2's methods depend upon — a semantic domain tagger equivalent to USAS, a validated effect lexicon equivalent to LIWC — does not exist in mature form for Urdu, and existing work on Urdu sentiment analysis has repeatedly been characterized in the computational linguistics literature as constrained by a

Critical Review of Social Sciences and Humanities Volume 6, Issue 1, 2026

scarcity of annotated resources (Khattak, Asghar, Saeed, et al., 2021). Any Layer L3 module built without addressing this gap will inherit the gap.

Third, Series Paper 1 documented without operationalizing a communicative pattern of central importance to Pakistani digital discourse: the routine mixing of Urdu vocabulary rendered in Roman script with English within a single utterance. This code-switching pattern is well studied in the broader South Asian computational linguistics literature for Hindi-English contexts (Bali, Sharma, Choudhury, & Vyas, 2014; Solorio et al., 2014; Molina et al., 2016) but has no equivalent annotated resource for Urdu-English crisis discourse. A pipeline that ingests only cleanly monolingual text will systematically fail on precisely the register in which much Pakistani digital self-disclosure occurs.

1.4 Research Objectives

- To specify the design, sourcing strategy, and ethical governance of a Multilingual Suicide Discourse Corpus (M-SDC) in Urdu and Punjabi, matched by a control corpus of equivalent scale and register.
- To specify an annotation scheme semantic domain tagging, code-switch tagging, and a culturally derived izzat-register lexicon capable of supporting the same class of quantitative analysis (Keynes, collocation, effect size) that Series Paper 2 applied to English.
- To pre-register, as directional hypotheses rather than reported findings, the predictions this design generates, so that the eventual results are falsifiable against a stated prior rather than assembled post hoc.
- To specify the reliability, validation, and researcher-wellbeing protocols required before any data collection involving native-language crisis text begins.

1.5 Research Questions

- What corpus construction design can produce a native-script Urdu/Punjabi Suicide Discourse Corpus of sufficient scale and provenance diversity to support Keynes, collocational, and semantic-domain analysis comparable to Series Paper 2?
- What annotation architecture semantic tagging, code-switch tagging, culturally specific lexicon development is required to make Urdu/Punjabi suicidal discourse computationally legible in the same terms as the English-language evidence base?
- Do the universal markers established in Series Paper 2 (pain vocabulary, first-person pronoun elevation, future-tense depletion, absolutist language) replicate in native-script Urdu/Punjabi discourse, and do the culturally specific markers of Series Paper 1 (izzat, honor-shame lexis, domestic entrapment metaphor) show comparable statistical robustness once measured at corpus scale?

2. BACKGROUND AND RELATED WORK

2.1 Multilingual and Low-Resource Mental Health NLP

Computational work on language and mental health has been overwhelmingly conducted on English-language social media data (Coppersmith, Dredze, & Harman, 2014; De Choudhury, Counts, & Horvitz, 2013; Gkotsis et al., 2017), a pattern that recent multilingual and hate-speech detection work has begun to identify as a structural limitation of the field rather than an incidental one (Gorai & Shaw, 2024). Urdu-language NLP more broadly has been characterized as resource-poor: reviews of Urdu sentiment analysis consistently report a shortage of annotated corpora, validated lexicons, and standardized preprocessing pipelines relative to English or even to other South Asian languages such as Hindi (Khattak, Asghar, Saeed, et al., 2021). This scarcity is not a

Critical Review of Social Sciences and Humanities Volume 6, Issue 1, 2026

peripheral inconvenience for the PFLSI programme; it is the central obstacle that Layer L3 of the PFLS pipeline must overcome, and no amount of architectural specification in Series Paper 3 substitutes for the underlying resource-building work. Within the PFLSI programme, a comparable English-language corpus study demonstrated that lexical and grammatical variables such as pain vocabulary, first-person reference, future-orientation, and absolutist language can be operationalized as statistically measurable indicators of suicidal crisis (Ishfaq & Sultan, 2026a).

2.2 Code-Switching as a Structural Feature, Not Noise

Work on Hindi-English code-mixed social media text — linguistically and computationally the closest well-studied analogue to Urdu-English digital discourse — has established that code-switching in South Asian online communication is pervasive, systematic, and resistant to naive language-identification approaches (Bali, Sharma, Choudhury, & Vyas, 2014; Sultan & Ishfaq, 2026). The CALCS shared-task series formalized word-level language identification as a distinct computational problem, introducing evaluation frameworks and annotation conventions — including tags for ambiguous and mixed-language tokens — that this paper adopts as the starting point for an Urdu-English scheme (Solorio et al., 2014; Molina et al., 2016). Series Paper 1's observation that Pakistani digital communication routinely mixes Roman-script Urdu vocabulary with English is therefore not an idiosyncratic finding requiring a bespoke solution; it is an instance of a well-characterized phenomenon for which partial computational tooling already exists and can be adapted rather than built from nothing.

2.3 Positioning Within the PFLSI Programme

This paper occupies a specific and bounded position in the programmer's roadmap. It does not propose a new theoretical framework the seven frameworks integrated in Series Paper 3 are treated here as the interpretive lens through which the new corpus will be read. It does not propose a new pipeline layer the corpus is built specifically to give empirical content to the already-specified Layer L3. And it does not report classifier performance that is reserved, per the roadmap set out in Series Paper 3, for Series Paper 5's transformer fine-tuning work, which depends on this paper's corpus as training data (Ishfaq et al., 2026). Series Paper 4 is, in the strictest sense, infrastructure: it exists so that later papers in the series have something real to compute on.

3. CORPUS DESIGN AND CONSTRUCTION METHODOLOGY

3.1 Overall Design Logic

The Multilingual Suicide Discourse Corpus (M-SDC) is designed to replicate, as closely as the source material allows, the multi-source logic of Series Paper 2's English-language SDC: heterogeneous provenance intended to test whether a linguistic signature is a property of suicidal discourse in general or an artefact of a single register or platform. Four source streams are specified, described in Section 3.2, together with a matched control corpus. All primary data collection is contingent on institutional ethics approval and formal data-sharing agreements; this paper specifies the design that such agreements would need to support, and treats data collection itself as not yet undertaken.

3.2 Data Sources

3.2.1 Crisis Helpline Transcripts (Urdu-Medium)

The primary and highest-priority source stream is anonymized transcripts of Urdu-medium crisis helpline interactions, obtained through a proposed institutional partnership with one or more Pakistani mental health helplines operating under existing ethical and clinical governance. This mirrors the highest-fidelity source stream in Series Paper 2's English SDC and is expected to

Critical Review of Social Sciences and Humanities Volume 6, Issue 1, 2026

provide the most direct linguistic evidence of acute suicidal crisis, subject to the specific consent, redaction, and data-sovereignty protocols described in Section 3.4.

3.2.2 Moderated Mental-Health Forums and Roman-Urdu Digital Communities

A second stream comprises publicly posted, terms-of-service-compliant text from Urdu- and Roman-Urdu-medium mental health forums and moderated social media communities analogous in function to the r/Suicide Watch and r/depression sources used in Series Paper 2. This stream is expected to be the primary source of code-switched material and will be the main testbed for the code-switch annotation scheme specified in Section 3.5.2.

3.2.3 Re-Digitized Original-Language Series Paper 1 Notes

Where the original Urdu- or Punjabi-language text of the Series Paper 1 note corpus (N=140) remains available in its source form, this stream proposes its re-transcription and inclusion, subject to renewed ethical clearance, as a validation subset: a bridge between the qualitative, thematically coded findings of Series Paper 1 and the quantitative, corpus-linguistic methods being extended to Urdu and Punjabi for the first time in the present design. This subset is intended to test, at the level of individual documents, whether the gendered and culturally specific patterns Series Paper 1 identified through manual coding are recoverable through automated Keynes and collocational methods.

3.2.4 Published First-Person Narratives

A fourth, smaller stream comprises published Urdu-language first-person narrative accounts of suicidal crisis and survival, sourced from print and digital publications with appropriate copyright and consent clearance, providing a register distinct from both clinical transcription and unmoderated social media.

3.2.5 Matched Control Corpus

A control corpus of equivalent target scale, matched by register, platform type, and approximate publication period, is specified using the same non-clinical Urdu/Punjabi social media, forum, and published-narrative sources, excluding any content flagged for suicide-related content during the sampling process. This mirrors the function of the 449,800-token Matched Control Corpus (MCC) in Series Paper 2.

3.3 Target Scale and Sampling Frame

The design target is a 500,000-token Suicide Discourse Corpus (Urdu/Punjabi) and a matched 500,000-token control corpus, a scale chosen to exceed Series Paper 2's English SDC (452,000 tokens) (Ishfaq & Sultan, 2026a) so that the smaller average word-to-token ratio of Urdu morphology does not leave the corpus statistically underpowered relative to its English counterpart. Table 1 specifies proposed source proportions.

Table 1

Proposed Source Composition of the Multilingual Suicide Discourse Corpus (Design Target)

Source Stream	Language / Script	Target Token Share	Target Tokens (approx.)
Crisis helpline transcripts	Urdu (Nastaliq / Urdu script)	35%	175,000
Moderated forums / Roman-Urdu social media	Roman-Urdu, Urdu-English code-switched	35%	175,000

Critical Review of Social Sciences and Humanities Volume 6, Issue 1, 2026

Source Stream	Language / Script	Target Token Share	Target Tokens (approx.)
Re-digitised Series Paper 1 notes	Urdu / Punjabi (Shahmukhi)	10%	50,000
Published first-person narratives	Urdu	10%	50,000
Punjabi (Shahmukhi) supplementary stream	Punjabi	10%	50,000

Note: figures are design targets for the sampling frame, not achieved collection totals. Final proportions will be revised once helpline and forum partnership agreements are finalized.

3.4 Ethical Sourcing and Governance

All primary data collection is contingent on institutional ethics board approval in addition to the approvals already secured for Series Papers 1 and 2, and on formal data-sharing and data-sovereignty agreements with any partner helpline or platform. Consistent with the ethical minimalism principle articulated in Series Paper 3, the design specifies: full anonymization and removal of identifying detail prior to any researcher access; storage of raw transcripts only for the minimum period required for annotation, with feature-level data retained thereafter; a prohibition on any secondary use of helpline-sourced material beyond the stated research purpose; and a translator/annotator wellbeing protocol, described in Section 6, given that annotators working directly with native-language crisis text face a qualitatively different exposure than annotators working with Keynes statistics computed over English text they did not personally translate.

3.5 Annotation Architecture

3.5.1 Semantic Domain Tagging

In the absence of a mature Urdu equivalent to the UCREL Semantic Analysis System (USAS) used in Series Paper 2 (Rayson et al., 2004; Ishfq & Sultan, 2026a), this design specifies a bilingual bridging approach: construction of an Urdu-to-USAS domain-mapping lexicon, built by bilingual annotators assigning USAS domain codes to high-frequency Urdu lemmas identified in pilot samples, followed by manual verification of domain assignment on a held-out sample. This produces an Urdu semantic-domain tagger that is domain-compatible with Series Paper 2's English analysis without requiring that a full independent Urdu semantic taxonomy be built from first principles.

3.5.2 Code-Switch Tagging

Each token in code-switched material is to be tagged for source language (Urdu, Punjabi, English, or ambiguous/mixed) following the token-level annotation conventions established in the CALCS shared-task tradition (Solorio et al., 2014; Molina et al., 2016), adapted for the Urdu-English pair. This tagging layer is what allows Layer L1 of the PFLS pipeline's proposed code-switching normalizer (Series Paper 3, Table 3) to be trained and evaluated rather than remaining an architectural placeholder.

3.5.3 The Izzat-Register Lexicon

Drawing on the qualitative thematic categories identified in Series Paper 1 — apologetic-relational discourse, izzat and family-obligation framing, coercion and forced-circumstance lexis — this design specifies the inductive construction of a culturally specific lexicon of honors-shame

Critical Review of Social Sciences and Humanities Volume 6, Issue 1, 2026

and relational-guilt vocabulary, built through iterative annotation of pilot samples by bilingual coders with cultural and clinical familiarity with the Pakistani context, and validated against the thematic categories already established in Series Paper 1 as a form of construct validity check.

3.6 Inter-Annotator Reliability

The design specifies dual independent coding of a minimum 15% overlap sample across all three annotation layers (semantic domain, code-switch, izzat-register), with Cohen's kappa (Cohen, 1960) computed per layer and a target threshold of $\kappa \geq 0.75$ before full-corpus annotation proceeds, consistent with the $\kappa = 0.81$ reliability achieved for thematic coding in Series Paper 1. Layers falling below threshold trigger a scheme-revision and re-training cycle rather than proceeding to full annotation, following standard practice in mixed-methods corpus construction (Landis & Koch, 1977; Teddlie & Yu, 2007).

3.7 Planned Analytic Procedure

Once annotation is complete, the design specifies replication of Series Paper 2's analytic pipeline on the new corpus: log-likelihood Keynes analysis (G^2) of the Urdu/Punjabi SDC against its matched control corpus; Mutual Information collocational analysis within a ±4–5 token window; chi-square comparison, with Bonferroni correction, (Ahmad, 2026) of semantic domain proportions between SDC and control; and Cohen's d effect sizes for the principal lexical variables identified in Series Paper 2 (pain/suffering vocabulary, first-person singular pronouns, future-tense forms, absolutist and negation markers) plus the new izzat-register variable specific to this corpus.

4. ANTICIPATED FINDINGS FRAMEWORK (PRE-REGISTERED HYPOTHESES)

Given that corpus construction is ongoing, empirical findings are not yet available. Accordingly, this section presents preregistered hypotheses rather than observed results. It instead pre-registers, as directional hypotheses derived from the convergent evidence of Series Papers 1 and 2, (Ishfaq et al., 2025; Ishfaq & Sultan, 2026a) the pattern of findings the M-SDC is expected to produce. Presenting hypotheses in this form — rather than waiting to state predictions only after data are in hand — allows the eventual analysis to be evaluated against a stated prior and reduces the risk that the programmer's cross-linguistic claims are constructed post hoc. Table 2 states each hypothesis and its theoretical anchor; Table 3 provides the shell of the results table that will be populated once analysis is complete, with all statistics marked as pending.

Table 2

Pre-Registered Hypotheses for the Multilingual Suicide Discourse Corpus

ID	Hypothesis	Theoretical Anchor	Basis in Prior Series Papers
H1	Pain/suffering vocabulary will show a large effect size ($d > 0.8$) in the Urdu/Punjabi SDC relative to its matched control, replicating the largest effect observed in Series Paper 2 ($d = 1.27$).	Shneidman (1993) psychache theory	S2 largest effect size across all variables
H2	First-person singular pronoun elevation will replicate at a medium-to-large effect size, consistent with S1's 78%/69% (female/male) note-level frequencies and S2's $d = 0.74$.	Pennebaker (2011) linguistic relativity	Converging S1 and S2 evidence

Critical Review of Social Sciences and Humanities Volume 6, Issue 1, 2026

ID	Hypothesis	Theoretical Anchor	Basis in Prior Series Papers
H3	Future-tense / future-oriented temporal reference will be depleted in the SDC relative to control, replicating S2's T1.3 domain reduction (4.47% to 1.89%).	Beck (1963, 1974) hopelessness construct	S2 domain-level finding
H4	Izzat-register and relational-guilt lexis, coded via the lexicon specified in 3.5.3, will show significantly higher keyness in female-authored SDC material than male-authored material, replicating S1's gendered chi-square finding (apologetic language, $\chi^2(1)=18.43$, $p<.001$).	Izzat / honour-shame discourse; Joiner (2005) IPTS extended culturally	S1 gendered thematic finding
H5	A code-switched analogue to the 'tired of existing' indirect-displacement collocation documented in S2 (MI = 7.2) will be identifiable in the Roman-Urdu / code-switched stream, though its specific lexical form is not predicted in advance.	O'Connor (2011) IMV model; indirect lexical suicidal displacement	S2 collocational finding
H6	Absolutist and negation language will show a medium effect size (d approx. 0.5–0.6) in the SDC relative to control, replicating S2's d = 0.58 for English absolutist words.	Al-Mosaiwi & Johnstone (2018); Shneidman (1993)	S2 effect size

Table 3

Results Table Shell — To Be Populated Following Corpus Completion

The following table reproduces the structure of Series Paper 2's Table 5 (Cumulative Statistical Evidence) so that Urdu/Punjabi results will be directly comparable, column for column, with the existing English-language findings. All statistic values are placeholders pending data collection and are marked accordingly; no numerical values in this table represent completed analysis.

Variable	Statistic	Value	p	Status
Pain/suffering lemmas: SDC vs Control (Urdu/Punjabi)	t-test, Cohen's d	[pending]	[pending]	Awaiting corpus completion
First-person singular pronouns: SDC vs Control	t-test, Cohen's d	[pending]	[pending]	Awaiting corpus completion
Future-tense depletion: SDC vs Control	t-test, Cohen's d	[pending]	[pending]	Awaiting corpus completion

Critical Review of Social Sciences and Humanities Volume 6, Issue 1, 2026

Variable	Statistic	Value	p	Status
Izzat-register lexis: gender comparison	$\chi^2(1)$	[pending]	[pending]	Awaiting corpus completion
Absolutist / negation language: SDC vs Control	t-test, Cohen's d	[pending]	[pending]	Awaiting corpus completion
Code-switch rate: SDC vs Control	Mann-Whitney U	[pending]	[pending]	Awaiting corpus completion

Note: this table is a placeholder shell, not a report of findings. It exists so that the analytic and reporting format is fixed in advance of, and independent of, the results it will eventually contain.

5. DISCUSSION

5.1 What Closing the Izzat Problem Would and Would Not Establish

If the hypotheses in Table 2 are supported once the M-SDC is complete, the programme will have established that the izzat-register and honors-shame findings of Series Paper 1 are not artefacts of a small, qualitatively coded sample but statistically robust features of Urdu/Punjabi suicidal discourse at corpus scale. This would matter for the same reason Series Paper 3 argued the izzat problem matters: it would mean that a detection system's blindness to culturally specific registers is a solvable resource gap rather than a fixed limitation of the underlying linguistic signal. It is equally important to state what corpus completion would not establish. A corpus, however large, is not a validated clinical instrument; replication of a lexical signature at scale is a necessary precursor to the transformer fine-tuning and prospective clinical validation work reserved for Series Papers 5 and later in the roadmap, not a substitute for it.

5.2 Implications for PFLS Pipeline Layer L3

The three components specified for Layer L3 in Series Paper 3 an Urdu/Punjabi sub-model, a metaphor detection module, and a code-switching normalizer each depend on a specific output of the present design. The semantic domain tagger specified in Section 3.5.1 provides the training signal for the sub-model; the izzat-register lexicon specified in Section 3.5.3 provides the seed vocabulary for the metaphor detection module's cultural source-domain coverage; and the code-switch tagging scheme specified in Section 3.5.2 provides the labelled data the normalizer requires. Until this corpus exists, Layer L3 remains, in the terms Series Paper 3 itself used, an architectural specification rather than a deployable component.

5.3 Anticipated Cross-Linguistic and Cross-Script Challenges

Several challenges are anticipated and are flagged here rather than deferred to a post hoc limitation's discussion. Punjabi in Pakistan is conventionally written in the Shanmukhi (Perso-Arabic) script, while the much larger body of digitally available Punjabi-language NLP resources assumes the Gurmukhi script used in Indian Punjab; any tooling reused from existing Punjabi NLP work will require script conversion or re-training. Roman-Urdu transliteration is highly non-standardized at the level of individual spelling choices, which complicates both language-identification and lexicon-matching and is expected to require a normalization step prior to semantic tagging (Ishfaq & Sultan, 2026b). Finally, dialectal and register variation within Urdu itself between the register typical of published narrative and the register typical of unmoderated social media may produce systematic differences in baseline lexical frequency that the matched

Critical Review of Social Sciences and Humanities Volume 6, Issue 1, 2026

control corpus design (Section 3.2.5) is intended to absorb, but which will need to be checked empirically rather than assumed away.

6. ETHICAL CONSIDERATIONS SPECIFIC TO THIS PAPER

Working directly with native-language crisis transcripts, rather than with pre-existing English-language datasets assembled by others, introduces ethical obligations beyond those already articulated in Series Paper 3's governance framework. First, annotators working with Urdu- and Punjabi-medium crisis text are working with linguistically and culturally proximal crisis narratives than annotators working with keyness tables computed over a corpus in a second or third language; the design specifies mandatory annotator wellbeing protocols, including capped daily exposure to raw crisis transcripts, access to debriefing support, and rotation off the annotation task on request without penalty. Second, the design commits to data-sovereignty terms in any helpline or platform partnership agreement that keep raw transcripts within Pakistani institutional custody wherever legally and technically feasible, consistent with the ethical minimalism principle's purpose-limitation requirement. Third, all illustrative examples used in any published output of this work will be composited, paraphrased, or drawn only from already-published Series Paper 1 material, following the safe-messaging practice already adopted in Series Paper 3 (Niederkrotenthaler et al., 2020).

7. LIMITATIONS OF THE PROPOSED DESIGN

- Institutional dependency: the highest-value data stream (crisis helpline transcripts) depends on partnership agreements that have not yet been secured at the time of writing; the design proceeds on the assumption that such agreements are achievable but cannot guarantee the proposed 35% source share if partnership negotiations do not succeed.
- Absence of a mature Urdu semantic tagger means the bilingual bridging approach in Section 3.5.1 is itself a novel methodological contribution requiring its own validation, rather than an off-the-shelf tool being applied.
- The re-digitization stream (Section 3.2.3) is contingent on the continued availability and re-consent ability of the original Series Paper 1 note materials, which may not be fully recoverable in original-language form for every note in the corpus.
- As a protocol paper, this contribution cannot report statistical power calculations grounded in observed variance; target corpus scale (Section 3.3) is set by analogy to Series Paper 2 rather than by a formal power analysis (Ishfaq & Sultan, 2026a), and may require revision once pilot data are available.

8. SERIES PROGRAMME ROADMAP UPDATE

This paper does not alter the PFLSI roadmap proposed by **Ishfaq et al. (2026)**; instead, it makes the anticipated Series Paper 5 on transformer fine-tuning and model validation methodologically achievable. Series Paper 5's proposed fine-tuning of XLM-Roberta and clinical BERT variants explicitly requires a multilingual training corpus; that corpus is the direct output of the present design (Ishfaq et al., 2026). Two further items are added to the roadmap as a consequence of the present paper's design work. First, a dedicated validation study reporting the reliability statistics (Section 3.6) and the completed results shell (Table 3) once corpus collection concludes, which may be published as a technical report preceding Series Paper 5 rather than folded into it. Second, given the anticipated script and dialect complications noted in Section 5.3, a short methodological note on Shanmukhi Punjabi NLP tooling is identified as a plausible standalone contribution to the broader field, independent of its role within the PFLSI programme.

9. CONCLUSION

Critical Review of Social Sciences and Humanities Volume 6, Issue 1, 2026

Ishfaq (2026) argued that language is among the earliest and most accessible signals of suicidal crisis and that a deployment-ready detection system must recognize that signal across languages and communicative registers. The present paper does not yet demonstrate that the PFLS pipeline can process Urdu and Punjabi as reliably as it can hear English; it specifies, in the detail an engineering and ethics review would require, how the evidence needed to make that demonstration will be built. The corpus described here does not yet exist. Its absence is, at present, the single largest gap between the PFLSI programmer's architectural ambition and its empirical foundation. Closing that gap carefully, with the ethical governance the material demands, and without overstating what a completed corpus alone can establish is the specific and bounded task this paper sets for itself and for the series paper that will follow it.

References

Ahmad, A. I. S. (2026). Language of influence: A corpus-based lexical analysis of psychological power strategies in Robert Greene's *The Laws of Human Nature*. In *Proceedings of the 2nd Riphah International Conference on Language, Literature, and Culture*.

Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clinical Psychological Science, 6(4), 529-542.

Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014). "I am borrowing ya mixing?" An analysis of English-Hindi code mixing in Facebook. Proceedings of the First Workshop on Computational Approaches to Code Switching, 116-126.

Beck, A. T. (1963). Thinking and depression: I. Idiosyncratic content and cognitive distortions. Archives of General Psychiatry, 9(4), 324-333.

Beck, A. T., Weissman, A., Lester, D., & Trexler, L. (1974). The measurement of pessimism: The Hopelessness Scale. Journal of Consulting and Clinical Psychology, 42(6), 861-865.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. Proceedings of the Workshop on Computational Linguistics and Clinical Psychology, 51-60.

Creswell, J. W. (2013). Qualitative inquiry and research design: Choosing among five approaches (3rd ed.). SAGE Publications.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 3267-3276.

Denzin, N. K., & Lincoln, Y. S. (2018). The SAGE handbook of qualitative research (5th ed.). SAGE Publications.

Gkotsis, G., Oellrich, A., Hubbard, T., Dobson, R., Liakata, M., Velupillai, S., & Dutta, R. (2017). Characterisation of mental health conditions in social media using informed deep learning. Scientific Reports, 7, 45141.

Gorai, A., & Shaw, S. (2024). Multilingual hate speech and suicidal ideation detection using large language models. Natural Language Engineering, 30(3), 578-601.

Hussain, N., Dubey, I., Saeed, W., & Patel, M. X. (2022). Suicide in Pakistan: Epidemiology, contributory factors, and opportunities for prevention. Journal of Affective Disorders, 312, 57-67.

Critical Review of Social Sciences and Humanities Volume 6, Issue 1, 2026

Ishfaq, A. (2025). Ethnolinguistic identity and cultural memory in the Dawoodi community. *Annual Methodological Archive Research Review, 3*(6), 185–208.

Ishfaq, A. (2026). *Thinking through language: Linguistic foundation and advanced academic writing*.

Ishfaq, A., & Bhatti, A. M. (2019). Language shift and imminent language death: A diachronic study of Dawoodi. *Harf-o-Sukhan, 3*, 1–14.

Ishfaq, A., & Bhatti, A. M. (2020). Lexical attrition and generational language competence in Dawoodi speakers. *Harf-o-Sukhan, 4*, 4–18.

Ishfaq, A., & Bhatti, A. M. (2021). From pidgin to creole to collapse: The evolutionary trajectory of Dawoodi language. *Jahan-e-Tahqeeq, 4*.

Ishfaq, A., & Bhatti, A. M. (2022). Linguistic hegemony and the silencing of Dawoodi: Power, stigma, and structural marginalization. *Jahan-e-Tahqeeq, 5*(4), 48–60.

Ishfaq, A., & Bhatti, A. M. (2023). Code-switching, borrowing, and linguistic dilution: Contact-induced change in Dawoodi. *Jahan-e-Tahqeeq, 6*(3), 577–592.

Ishfaq, A., & Bhatti, A. M. (2024). From heritage to liability: Language attitudes and identity reconstruction as drivers of obsolescence in the Dawoodi language. *Al-Mahdi Research Journal (MRJ), 5*(3), 1303–1336.

Ishfaq, A., & Sultan, N. (2024). Identification of different methodologies for treatment of autism in Urdu-speaking adolescents: An investigative report. *Contemporary Journal of Social Science Review, 2*(4), 1611–1618.

Ishfaq, A., & Sultan, N. (2025a). Cognitive control and executive function in advanced second-language writing. *Sareer-a-Khama, 4*(4).

Ishfaq, A., & Sultan, N. (2025b). Narrative and meaning in Surah Yūsuf: A critical hermeneutic analysis. *AL-HAYAT Research Journal (AHRJ), 2*(4), 11–23.

Ishfaq, A., & Sultan, N. (2025c). Trauma, resilience, and narrative healing: A psycho-hermeneutic reading of Surah Yūsuf. *AL-JAMEI Research Journal, 3*(1), 229–239.

Ishfaq, A., & Sultan, N. (2026a). Language of stress: A corpus-based study to detect early signs of suicide through lexical choice [PFLSI Series Paper 2]. Journal of Applied Linguistics and TESOL (JALT), 9(1), 862-880.

Ishfaq, A., & Sultan, N. (2026b). Modeling meaning in generative AI: A corpus-assisted discourse analysis of coherence, framing, and persuasion. *Pakistan Journal of Social Science Review, 5*(3), 1147–1164.

Ishfaq, A., Ahmad, S., & Sultan, N. (2025a). Decoding despair: A multidisciplinary psycho-forensic linguistic approach to suicide notes [PFLSI Series Paper 1]. Journal of Psychology, Health and Social Challenges, 3(2), 83-89.

Ishfaq, A., Ahmad, S., & Sultan, N. (2025b). Decoding despair: A multidisciplinary psycho-forensic linguistic approach to suicide notes. *Journal of Psychology, Health and Social Challenges, 3*(2), 83–89.

Ishfaq, A., Azim, M. U. (2025). Phono-semantics and translation: A cross-linguistic study of Urdu and Punjabi ideophones. *International Research Journal of Arts, Humanities and Social Sciences, 2*(3).

Ishfaq, A., Malik, A. H., & Sultan, N. (2025). Developing trauma-sensitive pedagogical practices for resilient learning in academia: A multidisciplinary approach of psycholinguistics and ELT. *Al Aasar, 2*(1), 171–189.

Ishfaq, A., Sultan, N., & Bhatti, A. M. (2026). The lexicon of crisis: A psycho-forensic linguistic architecture for real-time suicide risk detection integrating big data, corpus evidence, and theoretical frameworks [PFLSI Series Paper 3]. Annual Methodological Archive Research Review, 4(6).

Ishfaq, A., Sultan, N., & Healy, B. (2025). Turn-taking, politeness, and identity: A conversational study of *Speak Your Heart*. *Journal of Applied Linguistics and TESOL (JALT), 8*(3), 1567–1581.

Ishfaq, A., Sultan, N., Hassan, N., Aleem, F., & Maldonado, M. G. (2022). Elif Shafak's *Forty Rules of Love*: Contextual variation in adjectives. *International Online Journal of Language and Literature*.

Joiner, T. E. (2005). Why people die by suicide. Harvard University Press.

Khan, I. A., Khaled, F., & Ishfaq, A. (2025). Operation Bunyan un Marsoos: A critical analysis of human rights compliance—A study of the operation's adherence to human rights law and international humanitarian law. *Dialogue Social Science Review (DSSR), 3*(6), 173–184.

Khattak, A., Asghar, M. Z., Saeed, A., Hameed, I. A., Ahmad, S. A., & Hassan, M. (2021). A survey on sentiment analysis in Urdu: A resource-poor language. Egyptian Informatics Journal, 22(1), 53-74.

Kovecses, Z. (2015). Where metaphors come from: Reconsidering context in metaphor. Oxford University Press.

Lakoff, G., & Johnson, M. (1980). Metaphors we live by. University of Chicago Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174.

Menon, R. (2024). The narrative-crisis model: Understanding suicidal ideation through personal storytelling. Journal of Mental Health and Narratives, 12(3), 193-208.

Mishara, B. L., & Weisstub, D. N. (2016). The legal status of suicide: A global review. International Journal of Law and Psychiatry, 44, 54-74.

Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., & Solorio, T. (2016). Overview for the second shared task on language identification in code-switched data. Proceedings of the Second Workshop on Computational Approaches to Code Switching, 40-49.

Niederkrotenthaler, T., Stack, S., Till, B., Sinyor, M., Pirkis, J., Garcia, D., & Tran, U. S. (2020). Association of increased youth suicides in the United States with the release of 13 Reasons Why. JAMA Internal Medicine, 180(7), 1007-1014.

O'Connor, R. C. (2011). Towards an integrated motivational-volitional model of suicidal behaviour. In R. C. O'Connor, S. Platt, & J. Gordon (Eds.), International handbook of suicide prevention (pp. 181-198). Wiley-Blackwell.

Pennebaker, J. W. (2011). The secret life of pronouns: What our words say about us. Bloomsbury Press.

Rayson, P., Archer, D., Piao, S., & McEnery, T. (2004). The UCREL Semantic Analysis System. Proceedings of the LREC Workshop on Beyond Named Entity Recognition.

Critical Review of Social Sciences and Humanities Volume 6, Issue 1, 2026

Resnik, D. B. (2020). Research ethics: A philosophical guide to the responsible conduct of research (2nd ed.). Springer.

Shneidman, E. S. (1993). Suicide as psychache: A clinical approach to self-destructive behavior. Jason Aronson.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., & Fung, P. (2014). Overview for the first shared task on language identification in code-switched data. Proceedings of the First Workshop on Computational Approaches to Code Switching, 62-72.

Sultan, N., & Ishfaq, A. (2026). Instagram captions as identity performance: A multimodal discourse analysis. *AL-HAYAT Research Journal (AHRJ)*, *3*(2), 9–21.

Teddlie, C., & Yu, F. (2007). Mixed methods sampling: A typology with examples. Journal of Mixed Methods Research, 1(1), 77-100.

World Health Organization. (2023). Suicide worldwide in 2019: Global health estimates. WHO Press.