Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Light-Verb Innovation and Bilingual Predicate Creativity in Pashto-English: A Morpheme-Level Audit

ARSHAD Ali KHAN

arshad.khan@umt.edu.pk

University of Management and Technology

Amina Khalid

University of Management and Technology

Article

Keywords: bilingual compound predicates, light-verb innovation, predicate creativity, code-switching, 4-M Model, differential election, Pashto-English, matrix language

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10812879/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Bilingual predicate creativity can involve the distribution of lexical meaning and finite morphology between languages: an English-origin element provides event-related meaning, and a Pashto light verb or copula forms the bilingual predicate and encodes tense, aspect, agreement, and clause structure. This is a secondary morpheme-level audit of English-containing sentences from about eight hours of Pashto-dominant television interaction. The bounded interlinear appendix was screened to reconstruct 41 English-origin predicate candidates: 35 Core cases, with the host verb being Pashto kaw 'do/make' (29) or keg/shw 'become' (5) or the copula (1); five category-ambiguous Sensitivity cases; and one Excluded motion-verb construction. None of the 40 hosted candidates bears overt English tense or agreement morphology. The analysis distinguishes between observable surface form and syntactic category and borrowing status, since English bare forms are categorically ambiguous. The paper proposes a graded Differential Election Hypothesis (DEH) based on the existing low-competition account of embedded-language nonfinite verbs and matrix-language 'do' constructions.

1. Introduction

Lexical meaning and grammatical inflection may be from different languages when bilingual speakers put a predicate-like element from one language into a clause organised by another. In many bilingual varieties, a donor-language lexical element is used with a matrix-language light verb or verbaliser, and tense, aspect, mood, voice and agreement are in the matrix language. Pashto-English offers a clear example: English-origin elements are combined with Pashto kaw 'do/make', keg 'become', or a copula with finite morphology. The central empirical question is thus how light-verb innovation can account for bilingual predicate creativity and what it can tell us about the division of labour in the areas of lexical selection, argument-structure assembly, and grammatical encoding.

Previous accounts converge on an asymmetric pattern: embedded-language material supplies lexical meaning, while matrix-language resources build the finite predicate (Backus, 1996; Muysken, 2000; Myers-Scotton & Jake, 2014, 2017). Matrix-language 'do' constructions are especially relevant because they combine an embedded lexical element with matrix-language inflection. Constructional, borrowing, psycholinguistic, and computational approaches additionally make frequency, familiarity, semantic weight, and lexical competition viable explanations (Purmohammad et al., 2022; Salig et al., 2024; Tsoukala et al., 2021; Wohlgemuth, 2009).

There are three empirical problems that are not solved for Pashto-English. First, the finite-nonfinite distribution is not necessarily determined by published examples in an exhaustively defined dataset. Second, the forms play, divide, target, release, and collapse are morphologically bare but not uniquely identifiable as infinitives; the forms missing and connected may function as predicative adjectives or result-state participles. Third, the selection of kaw, keg, and the copula can be related to the event structure of the English lemma, which is an independently available Pashto construction, conventional

borrowing, or a combination of these. Thus, a surface preference for bare forms does not necessarily indicate a processing mechanism.

This study treats light-verb innovation as the empirical phenomenon and differential election as a graded, Pashto-specific operationalisation of Myers-Scotton and Jake's (2014) low-competition proposal. It asks how English-origin predicate candidates combine with kaw 'do/make', keg/shw 'become', and the copula; where tense, aspect, and agreement are realised; and what the hosts contribute to event and argument structure. Appendix-derived tokens are kept distinct from examples reproduced from earlier publications.

The graded Differential Election Hypothesis (DEH) retains the proposal that nonfinite embedded-language material can meet lexical-conceptual intentions without activating a competing finite system, while allowing candidates to differ in event-, argument-, and inflectional specification. The corpus can establish structural compatibility but cannot demonstrate processing economy or reduced production time; those claims require controlled behavioural, electrophysiological, or computational evidence.

Recent research clarifies the empirical gap that is the subject of this work. A large-scale comparison of bilingual compound verbs reveals that the semantic domains that allow for an other-language predicate are not uniform across contact communities, and that host distributions are not independent of event meaning, discourse domain, and local bilingual ecology (Balam et al., 2024). Meanwhile, a systematic review of bilingual code-switching research reveals significant variation in the operationalization of bilingual experience and code-switching behaviour, which further underscores the need for explicit sampling units, provenance, coding rules, and reproducible candidate definitions (Rayo et al., 2024). The supplied Pashto appendix is therefore considered as a bounded candidate inventory and the distinction between structural description and stronger claims about processing, community preference, or corpus-wide frequency is maintained.

The study poses four questions: (1) What predicate forms from English are found in the defined dataset with Pashto hosts? (2) What are the realisations of tense, aspect and agreement in the bilingual VP? (3) How do kaw, keg/shw, and the copula contribute to predicate creativity, event meaning, and argument-structure organisation in constructions containing English-origin elements? (4) What observations would distinguish differential election from construction frequency, accommodation, borrowing, and lexical competition?

The study contributes an auditable, category-neutral inventory of bilingual predicate formation. Within the bounded dataset, English-origin elements extend the lexical and event-semantic range of the Pashto VP, while Pashto hosts realise finite morphology and may contribute transitivity, change of state, and argument organisation. Predicate creativity is used conservatively to mean recurrent type diversity and recombination within this inventory, not experimentally demonstrated nonce productivity or community prevalence.

Core, Sensitivity, and Excluded cases are analysed separately because an English bare form is not uniquely infinitival and forms such as missing and connected may be adjectival or participial. Host totals are reported descriptively because opportunity denominators are unavailable. The graded DEH is therefore offered as a falsifiable extension: representational specification should explain residual host-choice or processing variation after frequency, borrowing status, morphological availability, and constructional familiarity are controlled.

2. Literature Review

2.1 The integration of embedded-language verbs in code-switching

One of the most common observations in the literature on code-switching is that embedded-language (EL) verbs do not tend to be autonomously inflected finite verbs in a matrix-language (ML) clause, but rather are incorporated in some way that leaves finite morphology to the ML (Backus, 1996; Muysken, 2000; Myers-Scotton & Jake, 2014, 2017). There are three patterns that are well attested. An EL nonfinite form can be used as a complement to an ML light verb, usually 'do/make'; an EL stem can be integrated by an ML verbaliser; or an EL root can be inflected by ML inflectional morphology. These strategies are asymmetric in that embedded lexical meaning is preserved while matrix-language resources build the finite predicate.

Myers-Scotton and Jake (2014) analyse matrix-language 'do' constructions as compromise structural strategies. The verbal inflections and matrix-language grammatical procedures are carried by a semantically reduced ML verb, and the critical lexical meaning is provided by an EL nonfinite form. According to them, this setup prevents a second congruence test between competing EL and ML finite specifications and can account for the apparent speed and ease of bilingual predicates. They do recognize, however, that English is diagnostically challenging because bare infinitives and many present-tense forms are the same, and that naturalistic distributions do not provide controlled time comparisons. Backus's (1996) Turkish-Dutch 'do' construction illustrates the relevant nativising strategy: the light verb furnishes ML predicate structure and inflection, while the Dutch element conveys the intended lexical meaning, as in (1).

(1) *Turkish–Dutch (*Backus, 1996, p. *238)*

… millet kijken yap ıyor

people watch-INF do-PROG-3SG

… everybody is watching.

The same configuration appears in Swedish–Persian, where a Swedish bare infinitive occurs to the left of the Persian auxiliary kardan ‘do’, which carries the inflectional morphology and the predicate-argument

structure, while the EL element supplies the semantics, as in (2) (Lotfabbadi, 2002, p. 111).

(2) *Swedish–Persian (*Lotfabbadi, 2002, p. *111)*

man-o besviken kard-i

me-ACC disappointed do.PST-2SG

You made me disappointed.

Muysken (2000) documents the same lexical template (V + 'do/make') in Indic and creole contexts, in which a leftmost EL verb carrying the semantics is conjoined with a host 'do' verb that is inflected at the formulator for the predicate-argument structure, as in the Sranan–Dutch and Tamil–English mixed verbs in (3)–(4).

(3) *Sranan / English / Dutch mixed verbs (as cited in* Muysken, 2000, p. *185)*

onti kare 'to hunt'; train kare 'to train'; bewijs kare 'to prove'

[EL verb] + kare 'do'

to hunt / to train / to prove

(4) *Tamil–English (*Annamalai, 1989, *as cited in* Muysken, 2000, p. *215)*

avan enne confuse pann-itten

he me confuse do-PST.1SG

He confused me.

In a second pattern, the EL stem is incorporated through a native verbaliser. German integrates loan verbs with -ier-en (verbaliser-INF), so that 'to arrange' becomes arrangieren (Myers-Scotton & Jake, 2017, p. 14); French verbs enter Brussels Dutch only once the stem is affixed with -er, as in (5), where the verbaliser is supplied by the ML and the EL verb is salient at the lexical-conceptual level (Treffers-Daller, 1994; cf. Muysken, 2000, p. 191).

(5) *French–Dutch (*Treffers-Daller, 1994*)*

bless-er-en, condamn-er-en, concurrenc-er-en

[Fr. stem]-VBZ-INF

to hurt / to condemn / to compete

Russian likewise nativises loan verbs with the suffix -ova- (klassificirovat' 'classify', telefonirovat' 'telephone'; Muysken, 2000, p. 195). A related strategy is morphological integration, in which an EL stem is flanked by ML affixes; Bokamba's Lingala–French data, cited by Backus (1996, p. 223), show a French infinitive stem preceded by Lingala prefixes and followed by tense–aspect suffixes. In English–Hindi (Siegel, 1987; as cited in Muysken, 2000, p. 208), everyday EL items such as sign and wait are activated preverbally and function as complements to Hindi kar 'do', which carries the inflection. Across these patterns the previous literature converges on a single generalisation: only nonfinite EL verbs are common, and EL finite verbs are rare (Backus, 1996; Muysken, 2000; Myers-Scotton & Jake, 2017). The present study takes this generalisation as its starting point and asks why it holds.

2.1.1 Recent evidence and unresolved classification problems

Two recent developments are particularly important for the present analysis. First, Balam et al. (2024) analyse 1,140 bilingual compound verbs from two contact Spanish communities and show that the semantic categories most open to other-language predicates differ by community. Their results caution against treating a frequent general-purpose light verb as a context-free grammatical preference: event semantics, discourse topic, community history, and the inventory of eligible alternatives must also be measured. For Pashto–English, this motivates token-level semantic coding and opportunity denominators for kaw, keg/shw, and copular environments.

Second, Treffers-Daller's (2025) Simple View argues that the borrowing-switching distinction should not be decided by token length or surface integration alone. Listedness—the extent to which an item or multiword unit belongs to the receiving community's lexical stock—is the central diagnostic, while recurrence, dictionary evidence, corpus association, and speaker knowledge provide operational evidence. This is directly relevant to English-origin items such as play, target, missing, connected, and brought up: morphosyntactic integration into a Pashto predicate does not by itself establish whether an item is a borrowing or an online switch. The neutral label predicate candidate is therefore theoretically preferable until community-level and audio-based diagnostics are available.

These classification problems intersect with a broader reproducibility concern. Rayo et al. (2024) show that code-switching studies vary markedly in their measurement of bilingual experience and switching context. For structural corpus work, the same lesson applies to corpus boundaries, candidate retrieval, speaker and programme provenance, coder independence, and the reporting of excluded or ambiguous cases. Reproducible analysis consequently requires a complete concordance, a distinction between Core and Sensitivity cases, independent coding and adjudication, and restriction of quantitative inference when exhaustive sampling or opportunity denominators are unavailable.

2.2 Bilingual language production and the mental lexicon

Research on bilingual production has long been concerned with the stages through which a preverbal message is converted into grammatical and phonological form and with how bilingual speakers access

one language while limiting interference from another (Garrett, 1975; Levelt, 1989; De Bot, 1992; Costa, 2005). Code-switching—alternation between languages within a single utterance—provides evidence that both languages remain active even when one is in use, yet the mechanisms underlying the production of code-switched utterances are still imperfectly understood.

One structure that foregrounds the role of grammatical category in production is the bilingual compound verb (BCV), in which a light verb from the ML combines with a nonfinite lexical verb from the donor language; the light verb 'supports' the construction while the lexical verb carries its semantic weight. Such [verb + verb] compounds have been reported for Tamil, Greek, Hindi, bilingual Navajo, and Spanish–English communities, among others. In Myers-Scotton's (1993) framework, the dominant language is the ML and the other the EL; models of the bilingual lexicon (Francis, 2005; Jiang, 2000; Kroll & Tokowicz, 2005) posit connections between words of the two languages at the levels of lexeme, lemma, and concept.

Wei (2002) develops a lemma-based account in which each lemma is tagged for a language and, when activated, supports the realisation of a lexeme and a set of morphosyntactic procedures in the formulator. The speaker's preverbal message activates language-specific semantic-pragmatic feature bundles that are mapped onto lemmas; activated lemmas then direct grammatical encoding at the functional level, followed by morphological and phonological encoding at the positional level. Crucially, the two languages do not contribute equally: one language has greater input into the resulting string (Myers-Scotton & Jake, 2000; Wei, 2001). Wei (2002) further argues that a switch is licensed when the EL and ML lemmas are congruent at the relevant level; where congruence fails, EL material surfaces as fixed expressions or EL islands that follow EL rather than ML morpheme order (Muysken, 2000; Khan, Khalid & Saleem, 2023).

These models share a layered conception of the abstract level, which carries the semantic, syntactic, pragmatic, morphological, and sociolinguistic information relevant to the speaker's intention (Wei, 2002). Following Myers-Scotton and Jake (2017, p. 9), the abstract lexical structure underlying a morpheme can be analysed at three levels: the lexical-conceptual level (semantic and pragmatic meaning), predicate-argument structure, and morphological realisation patterns. EL verbs are more puzzling, occurring rarely in some pairs and frequently in others. The analysis below offers a partial resolution by locating EL nonfinite verbs at the lexical-conceptual level and treating finite morphology as the province of the ML. The study considers only classic CS, in which the critical grammatical elements of bilingual constituents come from a single language, the ML.

Recent evidence requires the Pashto-English analysis to discriminate among mechanisms. Purmohammad, Vorwerg, and Abutalebi (2022) combined a conversational-corpus analysis of Persian-English bilingual compound verbs with a picture-word interference experiment. Their results suggest that an English verb can compete with the nominal constituent of a Persian complex predicate, directly linking naturalistic bilingual compound verbs to lexical selection. Tsoukala et al. (2021) showed computationally that auxiliary-switching asymmetries can emerge from differences in semantic weight

and distributional properties. Salig et al. (2024) further found that more frequent switch types produced smaller comprehension costs. Together, these studies make frequency, lexical competition, and constructional availability necessary alternatives to any account based on the number of election levels.

The Persian–English study also supplies a useful descriptive diagnostic that can be applied without assuming its experimental conclusion. For each bilingual compound verb, Purmohammad et al. (2022) asked whether a translation-equivalent monolingual Persian compound predicate exists, whether the English verb denotes the same event as the nominal constituent of that predicate, and whether the bilingual and monolingual constructions select the same support verb. Applied cautiously to Pashto–English, these questions distinguish three contributions that must not be collapsed: the event meaning associated with the English-origin element, the argument-structural or aspectual contribution of the Pashto host, and the tense–aspect–agreement morphology realised on that host. The present article uses this comparison to sharpen the description of the Pashto pattern; it does not assume that English verbs occupy nominal slots in every token or that lexical competition has been demonstrated for Pashto–English.

3. Theoretical Framework

3.1 The 4-M Model

The 4-M Model classifies morphemes by their syntactic role and by the way they are activated in production (Myers-Scotton & Jake, 2000). Its basic division is between conceptually activated and structurally assigned morphemes. Building on the content/system distinction of the Matrix Language Frame model, the 4-M Model separates morphemes into four types: content morphemes; early system morphemes; and two kinds of late system morpheme, bridges and outsiders. Content morphemes (prototypically nouns, verbs, descriptive adjectives, and free pronouns) are conceptually activated by the speaker's semantic-pragmatic intention. Early system morphemes (for example plural markers and determiners) are also conceptually activated, because they are triggered by their content-morpheme heads and add specificity to them. Late system morphemes differ from both: they are structurally assigned at the formulator.

3.2 The Differential Access Hypothesis

The Differential Access Hypothesis (DAH) specifies the level at which each morpheme type is elected. As Myers-Scotton (2002, p. 78) states, the morpheme types of the 4-M Model are differentially accessed in the abstract levels of production: content morphemes and early system morphemes are accessed at the level of the mental lexicon, whereas late system morphemes do not become salient until the level of the formulator. The DAH thus predicts a processing difference—not merely a categorial one—between conceptually activated and structurally assigned morphemes, and it is this prediction that the present study exploits.

3.3 Conceptually activated morphemes: content morphemes and early system morphemes

Under the DAH, content morphemes are activated at the mental lexicon by the speaker's preverbal intention and convey most of the semantic and pragmatic meaning of the clause; nouns are the most frequent content morphemes and receive the thematic roles assigned by verbs (Myers-Scotton & Jake, 2017). Other categories may also function as content morphemes; in (6), the EL preposition before assigns a temporal thematic role to its complement (Myers-Scotton, 1993, p. 124).

(6) *Swahili–English (*Myers-Scotton, 1993, p. *124)*

… u-let-e before kesho jioni.

2SG-bring-SBJV before tomorrow evening

… you should bring it before tomorrow evening.

3.4 Late system morphemes

Late system morphemes are elected at the formulator and are structurally assigned. They are of two kinds. Bridge system morphemes build larger constituents within a phrase and keep them well formed in the relevant language; the possessive of in English is a typical bridge. Outsider system morphemes depend on information outside their immediate constituent—subject–verb agreement is the clearest case. Because late system morphemes are responsible for the morphosyntactic architecture of the clause, the model predicts that any element requiring them will be elected, in part, at the formulator rather than at the lexical-conceptual level (Myers-Scotton & Jake, 2009).

3.5 The Differential Election Hypothesis for embedded predicate elements

The graded DEH is a Pashto-focused operational extension of Myers-Scotton and Jake's (2014) account, not an alternative to it. It proposes that English-origin material may contribute lexical-conceptual and event-related information while the Pashto host supplies finite specifications, and that coordination demands should increase when both languages specify the same representational domain.

Four assumptions guide the analysis: lexical content and structural morphology are distinguishable; predicate candidates may vary in event-, argument-, and inflectional specification; Pashto supplies the finite frame in the target clauses; and overlapping cross-language specifications should increase competition. The latter two claims about graded representation and competition are testable hypotheses rather than findings established by this corpus.

Accordingly, the corpus tests only whether the distribution is structurally compatible with the proposal. The incremental prediction is that representational specification should explain host choice or processing after frequency, borrowing, constructional familiarity, and morphological availability are controlled. Figure 1 summarises the proposed division of labour; it does not represent measured processing time.

Figure 1 translates Myers-Scotton and Jake's production architecture into the analytical division of labour examined here. The communicative intention triggers the lexical-conceptual material and a matrix-language frame; the English-origin candidate provides the lexical meaning and can impose restrictions on event or argument structure, while the Pashto construction provides the available predicate template and overt finite morphology. These specifications are coordinated at the formulator and are realised as an English-origin element hosted by kaw, keg/shw or a copula. The figure is a theoretically informed interpretation of the attested distribution, not evidence of processing time or competition; such claims would need independent experimental measures (Myers-Scotton & Jake, 2014; Purmohammad et al., 2022).

The updated DEH makes four discriminating predictions, each grounded in earlier work but formulated here as a testable Pashto-English claim. Prediction 1 (P1: morphological locus) states that English-origin predicate candidates in Pashto matrix clauses should normally lack overt English tense and agreement morphology; finite exponents should instead appear on the Pashto host. This prediction follows the Matrix Language Frame/4-M distinction between embedded-language content material and matrix-language morphosyntactic framing, as well as cross-linguistic accounts of light-verb accommodation (Myers-Scotton & Jake, 2000, 2014; Wohlgemuth, 2009). Prediction 2 (P2: host selection) states that the choice among kaw, keg/shw, and the copula should be systematically related to independently coded event structure and argument structure rather than merely to the surface form of the English-origin item. Prior constructional and bilingual compound-verb research supports treating light-verb choice as structurally and semantically consequential rather than as an empty insertion mechanism (Goldberg, 1995; Purmohammad et al., 2022). Prediction 3 (P3: residual explanatory value) states that, if differential election contributes beyond entrenchment, DEH-relevant variables should predict candidate form or host choice after controlling for lemma frequency, borrowing status, speaker, programme, and construction frequency. This prediction extends earlier accounts of differential morpheme activation and election by requiring the proposal to explain variation beyond borrowing and constructional familiarity (Myers-Scotton & Jake, 2009, 2017; Wohlgemuth, 2009). Prediction 4 (P4: processing cost) states that, in controlled production, stronger competition between English-origin and Pashto specifications should produce longer latencies, more disfluency or repair, or stronger electrophysiological responses, consistent with experimental work on bilingual compound-verb processing (Purmohammad et al., 2022). A productive class of autonomously English-finite verbs in otherwise Pashto matrix clauses would challenge P1; host choice unrelated to event or argument structure would challenge P2; and failure of DEH variables to explain residual variation beyond frequency and construction would challenge P3. P4 requires experimental evidence and cannot be evaluated from the present corpus alone. These predictions operationalise the proposal without treating the observed corpus distribution as direct

evidence of processing. The present audit evaluates P1 directly and P2 descriptively; P3 and P4 are prospective predictions requiring source-linked or experimental evidence.

4. Methodology

4.1 Corpus and sampling

This research was conducted in Pakistan. It is a secondary morpheme-level audit of Pashto-dominant AVT Khyber public-broadcast material accessed online for a broader study of Pashto-English bilingual speech. Approximately twelve hours from 35 clips were screened; 25 clips (about eight hours) were retained. Only the public-broadcast portion of the archive is analysed here, and no privately elicited interview data are included. Because English-containing sentences rather than complete programmes were transcribed, the retained interlinear appendix—not the full audiovisual corpus—is the bounded analytical dataset. Programme, topic, participant composition, and duration details are presented in Table 1.

The programme, topic, participant composition, and duration of the 25 retained clips are summarized in Table 1.

Table 1

Programme, topic, participants, and duration of the retained clips

Clip	Programme	Topic	Participants	Min.
1	Khyber Club, Ep. 1	Computers	M: 2; F: 2	14
2	Khyber Club, Ep. 11	Electronic/print media and its influence	M: 4; F: 1	14
3	Khyber Club, Ep. 16	Tuition and education	M: 2; F: 1	14
4	Khyber Club, Ep. 16	The use of tuition in education	M: 2; F: 1	12
5	Khyber Club, Ep. 14	Peace, prosperity and development	M: 3	14
6	Khyber Club, Ep. 11	Electronic media	M: 2; F: 1	14
7	Khyber Club, Ep. 13	Electronic media	M: 2; F: 1	14
8	Khyber News	My Right (women's issues)	F: 4	8
9	Khyber News	My Right (women's issues)	F: 2	8
10	Khyber News	My Right (women's issues)	F: 2	14
11	Khyber News	Music (Ismail)	M: 2; F: 2	19
12	Khyber News	Personal political views	M: 2; F: 2	14
13	Khyber News	Personal political views	M: 2; F: 2	14
14	Khyber News	Personal political views	M: 2; F: 2	14
15	Nun Sahar	Education and progress	M: 2	52
16	Nun Sahar	Self and reality	M: 1; F: 1	14
17	Nun Sahar	Self and reality	M: 1; F: 1	14
18	Khyber Club	Politics: caretaker prime minister	M: 4	14
19	Nun Sahar	Unemployment	M: 4	9
20	Nun Sahar	Unemployment	M: 4	14
21	Nun Sahar Classics	Motorway issues	M: 1; F: 1	39
22	Nun Sahar Classics	Islamabad police	M: 1; F: 1	27
23	Khpal Etwar	Doctor's opinion	M: 1; F: 2	27
24	Khpal Etwar	Doctor's opinion	M: 1; F: 2	21
25	Khpal Etwar	Female education	F: 3	14

4.2 Selection procedure

The purposive sampling targeted publicly accessible programmes in intelligible Yousafzai Pashto that contained spontaneous English-origin material across political, educational, cultural, and social topics. Mixing-rich selection supports structural description but not prevalence estimates because Pashto-only sentences and complete programme transcripts were not retained.

4.3 Transcription and coding

English-containing sentences were Romanised and processed in Toolbox on aligned morpheme-segmentation, gloss, and free-translation tiers. The bilingual clause is the structural unit. The predicate subset was coded for English-origin form, Pashto host, morphology locus, syntactic function, provenance, analytical status, and verification state. Stable IDs APPX-001-APPX-041 link the manuscript to the supplementary workbook. A bare English form was not automatically classified as an infinitive, and timestamps were not treated as an explanatory variable.

4.3.1 Reconstruction and audit of Appendix A

The 1,147-paragraph interlinear appendix was screened as a bounded archive. The re-audit reconstructed 41 strings: 40 English-origin elements hosted by kaw, keg/shw, or a Pashto copula, plus prepare ra-zam 'come prepared', excluded because ra-zam is a lexical motion verb. The original total of 40 hosted constructions is therefore consistent with the 41-item screening inventory. All strings were checked against the retained interlinear text; high-confidence source mappings are recorded where available.

Applying the category-neutral rules yielded 35 Core cases, five Sensitivity cases, and one Excluded case. A boundary scan also logged six overt NP-complement sequences with kaw (study, analysis, table talk, discussion twice, and calls); these are supplementary screening exclusions, not additional predicate candidates.

4.3.2 Candidate definition and eligibility rules

A predicate candidate is an English-origin element contributing action, event, state, or result meaning within a Pashto predicate. Core cases combine such an element with a Pashto light verb or copula bearing visible clause-level morphology. Category-ambiguous adjectival, participial, result-state, or multiword forms enter the Sensitivity set; non-target hosts are Excluded. Overt nominal complements diagnosed through determiner/possessive embedding, plural or derivational morphology, or compound-noun structure are excluded. This procedure avoids deciding category, finiteness, or borrowing status from English orthography alone.

4.3.3 Borrowing–switching diagnostics

Borrowing status is unresolved because the retained archive lacks sufficient recurrence, speaker, phonological, dictionary, and community-judgment evidence. The main result concerns overt morphology locus and does not depend on a categorical loan-versus-switch decision.

4.4 Coding reliability and audit status

No independent coder sheets or adjudication log survive, so no reliability coefficient is claimed. Transparency is supported by prespecified eligibility rules, a complete 41-item decision log, six documented boundary exclusions, and separate Core and Sensitivity analyses. Supplementary Data File S2 includes operational definitions and a blank future recoding template.

4.5 Analytical framework and statistical procedures

Analysis combined descriptive counts, morpheme-level interpretation of meaning and morphology, and constrained comparison of DEH, lexical-competition, constructional, and borrowing accounts. Translation equivalence, event correspondence, and support-verb correspondence were considered only when the archive supported them (Purmohammad et al., 2022). No goodness-of-fit test was used because a .50 or uniform baseline is unjustified; inferential host-choice modelling would require exhaustive sampling, opportunity denominators, and speaker/programme controls. Table 2 aligns the four research questions with their evidence bases, coding variables, and analytical procedures. Accordingly, the present audit evaluates P1 directly and P2 descriptively, whereas P3 and P4 are stated prospectively and are not tested with the retained archive.

Table 2
Alignment of research questions, evidence, coding, and analysis

Research question	Evidence base	Principal coding	Analytical treatment
RQ1: Surface forms and hosts	41 candidates (40 hosted)	Surface morphology; host; analytical status	Counts and percentages, with Core and Sensitivity totals separated
RQ2: Locus of TAM, agreement, and voice	40 hosted candidates	Locus of overt English/Pashto morphology	Morpheme-level description; no unmotivated null-hypothesis test
RQ3: Semantics, argument structure, and host	Core set plus Sensitivity cases	Event/result meaning; syntactic function; argument-structural diagnostics	Qualitative comparison across kaw, keg/shw, and copular hosts
RQ4: DEH versus alternative accounts	Audited distribution plus published comparative cases	DEH predictions; construction-frequency, borrowing, and accommodation diagnostics	Constrained theory comparison; processing claims treated as hypotheses

Host counts describe this appendix only. Grammatical preference would require denominators for eligible transitive, change-of-state, and copular contexts, including Pashto-only controls; therefore totals are not treated as selection probabilities.

The source container is the broadcast recording, the analytical unit is the bilingual clause, and the quantitative unit is the appendix-derived candidate. The archive is complete for the retained English-containing appendix but not for all speech in the eight broadcast hours. Claims are consequently limited to the documented constructions rather than broadcast-wide prevalence.

4.6 Ethical considerations

This research was conducted in Pakistan. Before data collection, the research plan was reviewed and approved through departmental academic oversight by the Departmental Graduate Committee (DGC), Department of Linguistics and Communications, School of Liberal Arts, University of Management and Technology, Lahore, Pakistan. The DGC approval did not carry a separate institutional research-ethics/IRB reference number or formal exemption letter. The present article is a secondary linguistic analysis solely of publicly broadcast AVT Khyber programmes and involved no recruitment, private elicitation, intervention, or new interaction with individuals. No direct personal identifiers are included. The research was performed in accordance with the University of Management and Technology's applicable research policies and relevant Pakistani requirements for academic research, copyright, and attribution. The Declaration of Helsinki was not applicable because this was not medical research and no human participants were recruited for the present study.

5. Results and Analysis

5.1 Distribution of English-Origin Predicate Candidates

In the original structural coding, 40 items were coded as bilingual compound predicates. The bounded interlinear appendix was screened using the present category-neutral screening procedure, resulting in 41 candidate strings. Thirty-five meet the conservative Core criterion: an English-origin lexical predicate is combined with a Pashto light verb or copula that bears the visible clause-level morphology. Five additional candidates (balance, missing, brought up, connected, and clear) are kept only for Sensitivity analysis because they allow adjectival, result-state, nominal, or multiword-participial analyses. The motion verb ra-zam is used with Prepare, which is Excluded because it is used with the motion verb, not the target light-verb/copular construction. The 40 hosted candidates are the same as the original construction count, but the present study revises their categorical analysis. There is no overt English tense or agreement morphology on any of the 40 hosted English-origin elements. The distribution of the 41 reconstructed candidates by analytical status and Pashto host is presented in Table 3.

Table 3
Audited English-origin predicate candidates by analytical status and Pashto host

Analytical status	kaw do/make	keg/shw become	Copula be	Non-target	Total
Core predicate candidates	29	5	1	0	35
Sensitivity candidates	0	3	2	0	5
Excluded candidates	0	0	0	1	1
All hosted candidates	29	8	3	0	40
All reconstructed candidates	29	8	3	1	41

Note. The primary analysis is based on core counts, with Sensitivity cases being analyzed separately. The non-target column includes prepare ra-zam 'come prepared', which is not included in the light-verb/copular analysis. All segmentations and glosses are still to be verified against the recordings (which are not available) and the original Toolbox project.

Table 4 presents the complete 41-item screening inventory so that readers can inspect the empirical boundary of the analysis directly. The Status column distinguishes the 35 Core cases from the five category-ambiguous Sensitivity cases and the single Excluded non-target construction. This token-level display makes the reported host totals and the Core/Sensitivity analysis fully traceable.

Table 4
Complete screening inventory of 41 English-origin predicate candidates

Audit ID	English-origin form	Surface form	Pashto host	Status
APPX-001	collapse	bare	keg/shw 'become'	Core
APPX-002	play	bare	kaw 'do/make'	Core
APPX-003	discuss	bare	kaw 'do/make'	Core
APPX-004	start	bare	kaw 'do/make'	Core
APPX-005	share	bare	kaw 'do/make'	Core
APPX-006	apply	bare	copula	Core
APPX-007	target	bare	kaw 'do/make'	Core
APPX-008	play	bare	kaw 'do/make'	Core
APPX-009	share	bare	kaw 'do/make'	Core
APPX-010	exaggerate	bare	kaw 'do/make'	Core
APPX-011	start	bare	kaw 'do/make'	Core
APPX-012	call	bare	kaw 'do/make'	Core
APPX-013	balance	bare/adjectival	keg/shw 'become'	Sensitivity
APPX-014	provide	bare	kaw 'do/make'	Core
APPX-015	portray	bare	kaw 'do/make'	Core
APPX-016	add	bare	kaw 'do/make'	Core
APPX-017	link	bare	kaw 'do/make'	Core
APPX-018	depend	bare	kaw 'do/make'	Core
APPX-019	divert	bare	kaw 'do/make'	Core
APPX-020	arrange	bare	kaw 'do/make'	Core
APPX-021	missing	-ing/adjectival	copula	Sensitivity
APPX-022	forward	bare	kaw 'do/make'	Core
APPX-023	provide	bare	kaw 'do/make'	Core
APPX-024	computerize	bare	keg/shw 'become'	Core
APPX-025	show	bare	kaw 'do/make'	Core
APPX-026	start	bare	kaw 'do/make'	Core
APPX-027	appreciate	bare	kaw 'do/make'	Core

Audit ID	English-origin form	Surface form	Pashto host	Status
APPX-028	prepare	bare	ra-zam 'come'	Excluded
APPX-029	divide	bare	kaw 'do/make'	Core
APPX-030	brought up	-ed multiword	keg/shw 'become'	Sensitivity
APPX-031	conclude	bare	kaw 'do/make'	Core
APPX-032	add	bare	kaw 'do/make'	Core
APPX-033	decide	bare	keg/shw 'become'	Core
APPX-034	use	bare	kaw 'do/make'	Core
APPX-035	release	bare	keg/shw 'become'	Core
APPX-036	use	bare	keg/shw 'become'	Core
APPX-037	connected	-ed/adjectival	copula	Sensitivity
APPX-038	clear	bare/adjectival	keg/shw 'become'	Sensitivity
APPX-039	attach	bare	kaw 'do/make'	Core
APPX-040	support	bare	kaw 'do/make'	Core
APPX-041	avail	bare	kaw 'do/make'	Core

Note. Each row represents one screened token rather than one unique lemma. APPX-001–APPX-041 are stable audit identifiers. Core = primary predicate dataset; Sensitivity = category-ambiguous hosted case; Excluded = screened non-target construction. All 40 hosted cases lack overt English tense or agreement morphology.

Descriptive uncertainty and limits of inference:

No null-hypothesis tests are reported for the reconstructed appendix distribution. A .50 finite/nonfinite null and uniform host/category nulls have no independent linguistic motivation, and the tokens are clustered by speaker, programme, lemma, and source. Inferential modelling requires programme and speaker mapping, exhaustive opportunity denominators, and sufficient observations to account for clustering.

5.1.1 Sensitivity and robustness of the distribution

The principal morphology-locus result is the same for all analytical strata. Overt English tense or agreement morphology occurs in 0 of 35 Core cases and remains absent in all five Sensitivity cases, yielding 0 of 40 hosted candidates overall. Host proportions are more classification-sensitive: kaw accounts for 29/35 Core cases (82.9%) but 29/40 hosted cases (72.5%) after the five ambiguous items are restored; keg/shw rises from 5/35 (14.3%) to 8/40 (20.0%), and copular hosting rises from 1/35

(2.9%) to 3/40 (7.5%). The surface division of finite labour is therefore robust to category uncertainty, while the exact host distribution is partially dependent on the classification of adjectival, participial and multiword cases. The analysis thus considers 0/40 overt English finiteness to be the most definitive descriptive result, and the 29/5/1 Core host profile to be a more tentative constructional distribution.

5.2 The transitive light verb kaw ‘do/make’

In the kaw construction, an English-origin predicate candidate combines with a Pashto light verb inflected for tense, aspect, and subject–verb agreement. The English element provides lexical meaning and can restrict the event or argument structure, and kaw provides the overt finite morphology. In (7), the attested English-origin form play is hosted by kaw, which has third-person singular present imperfective morphology.

(7) *Pashto–English*

Media [IP [NP negative role] hu [V play kaw-i]].

media.NOM negative role indeed play do.PRS.IPFV-3SG

The media indeed plays a negative role.

Example (7) provides a clear attested light-verb pattern: play identifies the lexical event, while kaw-i integrates that event into a Pashto clause and realises present imperfective and third-person singular morphology. Because English play has the same surface shape in infinitival and finite non-third-person contexts, it is analysed as a bare-form candidate rather than categorically as an infinitive. It is consistent with DEH prediction P1, but it is not independent evidence for differential election, since matrix-language, borrowing, and constructional accounts predict the same surface division.

From the tokens alone, it is not possible to determine whether these English-origin items replace Pashto counterparts in online production. Previous studies suggested some counterpart comparisons for some constructions, such as play kaw- and ada kaw-, divide kaw- and taqsim kaw-, and use kaw- and istimalaw-. Such comparisons are used to motivate a translation-equivalence diagnostic, but they do not in themselves imply identical lexical category, event structure, argument selection, or processing competition. As in Purmohammad et al. (2022), such claims must be backed up by independently verified monolingual predicates and confirmation that the proposed counterpart is in the same constructional slot.

(8) *Pashto–English*

che [VP [PP dwa groups ke] mong [V divide kr-o]].

COMP two groups in.OBL 1PL.NOM divide do.PRS-1PL

[That] we have divided [it] into two groups.

In (8), the morphology of divide is bare, while the morphology of kr-o is visible. The interlinear source glosses the host as present first plural, and the free translation uses an English perfect. The article does not assume a perfective value, as the recording and original Toolbox analysis are not available. The example nevertheless securely locates the overt person marking on the Pashto host rather than on divide.

(9) *Pashto–English*

pa de ke monga [VP [NP youth] [V target ka-wu]].

at DEM.PROX in.OBL 1PL.NOM youth target do.PRS.IPFV-1PL

In this, we target the youth.

In (9), target is an attested bare English-origin candidate, and ka-wu has present imperfective and first-person plural morphology. In isolation, it can be used in either nominal or verbal lexical-category analyses, but in the retained annotation it is in the lexical position of the bilingual predicate. The category-neutral Core classification therefore rests on constructional function and morphology locus, not on calling target an English infinitive. This is precisely the kind of cross-category uncertainty highlighted by Purmohammad et al. (2022): apparent translation equivalence does not guarantee that competitors belong to the same lexical category. Independent audio and monolingual-counterpart evidence would still be required to determine borrowing status and lexical-selection competition.

(10) *Pashto–English*

che [VP [NP youth] [NP hpala study] [V na kaw-i]].

COMP youth own study not do.PRS.IPFV-3SG

… that the youth are not doing their study.

Example (10) documents the lower boundary of the candidate definition rather than an additional Sensitivity token. Study is embedded in the possessive NP hpala study 'their study' and is therefore treated as an overt nominal complement of kaw, not as an English-origin predicate candidate. The same rule excludes analysis, table talk, two tokens of discussion, and calls when the supplied constituent annotation and overt morphology identify them as nominal complements. Reporting these exclusions prevents a coarse [English item + kaw + TAM] search from being mistaken for the theoretically defined predicate inventory.

5.3 The intransitive light verb keg ‘become’

Intransitive and resultative predicates are built on keg/shw 'become', which takes a single argument and, like kaw, carries the finite morphology. In (11), release conveys lexical event or result-state meaning, while shw- is used to construct an intransitive change-of-state construction and to express third-person

singular perfective morphology. The host thus does not only contribute inflection, but also plays a role in the structure of the event and the realisation of arguments.

(11) *Pashto–English*

Halaq way-i che [cassette release shw-a].

people say.PRS-3PL that.COMP cassette release become.PRS.PFV-3SG

People say that the cassette has been released.

In (11), release + shw-a is an inchoative or resultative predicate. The English element provides a result-state meaning, and Pashto become realises perfective morphology. The key issue is whether release is a bare verb, a category-neutral root, or a resultative predicate, and the current data do not allow a decision to be made on the basis of orthography alone.

The same surface structure is found in (12), where collapse contributes change-of-state meaning, and the intransitive host sh- licenses the single-argument structure and bears the visible tense, aspect, and agreement morphology. Combined with release in (11), this is the most transparent distributional test of P2: host choice is not just bare morphology, but also transitivity and result-state semantics. The pair thus elicits independently coded comparisons between event classes, but not cross-category lexical competition or a processing-cost advantage over a finite English alternative (Purmohammad et al., 2022).

(12) Pashto–English

Nu pura [system collapse sh-i].

then complete system collapse become.PRS.PFV-3SG

… then the whole system collapsed.

5.4 Copular hosts and EL islands

One additional resultative shw construction and three copular constructions are instructive. In (13), the English-origin sequence brought up combines with wa-sh-i. Because brought may retain English participial morphology, up may remain an English-selected particle, and sh- supplies the resultative host, the sequence is retained as a keg/shw Sensitivity case rather than classified as an unambiguous bare English verb or ordinary copular predicate. Examples (14)-(16) then illustrate the Core and Sensitivity boundaries among candidates hosted by the Pashto copula.

(13) Pashto–English

Che d aghwi hə [brought up wa-sh-i].

that.COMP of 3PL.DEM well brought up PFV-become.PRS-3PL

… that they should be brought up well.

Example (13) contains the multiword English unit brought up. It is analytically distinct from a general verb-particle category because brought may retain English participial morphology, up may remain a selected particle, and the sequence may constitute an EL island. Its English-internal structure therefore makes it theoretically different from the attested bare-form construction play kaw-i.

In (14) the EL verb apply is hosted by the Pashto copula marked for future and imperfective; the EL form supplies only the semantic-pragmatic content, while the projection for tense and aspect is realised on the copula at the level of predicate-argument structure (Wei, 2002, p. 704l; Khan et al., 2023).

(14) Pashto–English

də Islam toul golden rule ba halta [apply wi].

of Islam all golden rule FUT there apply COP.PRS.IPFV.3PL

All the golden rules of Islam will be applied there.

Finally, participial-looking forms with adjectival or result-state readings are hosted by the copula, as with missing in (15) and connected in (16). Their surface form does not settle their category. They therefore belong to the Sensitivity analysis and cannot independently support a claim about English verbal infinitives.

(15) *Pashto–English*

… che Ajab Khan da yaw dwa-o program-uno na [missing wo].

… that.COMP Ajab Khan of one two-OBL programme-PL.OBL from missing COP.PST.IPFV.3SG

… that Ajab Khan was missing from one or two programmes.

In (15), missing is naturally analysed as a predicative adjective or participial adjective selected by the copula. It is therefore separated from eventive light-verb predicates in the core analysis and included only in a sensitivity analysis of broadly verbal-origin predicates.

(16) *Pashto–English (*Khan & Muysken, 2014*)*

Taso dagha halaq-o sra [connected yain].

2PL those.DEM people-OBL with connected COP.PRS.IPFV.2PL

You are connected with those people.

Connected in (16) is similarly compatible with a result-state adjective or passive participle. The Pashto copula carries agreement, but this does not by itself establish that connected is an inserted lexical verb. The with-phrase is treated as a diagnostic of potentially retained English lexical structure, although the present token alone cannot decide between the competing category analyses.

Taken together, the attested hosts support one robust descriptive generalisation: overt finite morphology is realised on the Pashto light verb or copula rather than on the English-origin candidate. The English element contributes lexical-conceptual content and may also constrain event, result-state, or argument structure. The distribution is therefore compatible with a division of labour between an English-origin lexical contribution and Pashto finite morphology, while the degree of EL structural specification remains an empirical question rather than a categorical lexical-conceptual-only fact.

5.5 Within-lemma host alternation: use with kaw and keg/shw

Most candidates are found on only one host, and lemma identity and host choice are not generally separable. The two tokens of use provide a useful internal comparison. In (17a), use combines with transitive kaw; in (17b), the same English-origin form combines with intransitive/resultative keg. The contrast does not establish a categorical semantic rule, but it shows that host choice is not mechanically fixed by the English lemma alone.

(17a) Pashto–English

da har sok [use kaw-i].

DEM.PROX every person use do.PRS.IPFV-3SG

It is used by everybody.

(17b) Pashto–English

da computer [all over the world] [use keg-i].

DEM.PROX computer all over the world use become.PRS.IPFV-3SG

The computer is used all over the world.

The contrast is theoretically informative because Pashto itself distinguishes the transitive light verb kaw from the intransitive or change-of-state host keg/shw (Babrakzai, 1999). The two tokens are thus compatible with P2: the Pashto construction contributes to the construal of transitivity and event even when the English-origin lemma is the same. Meanwhile, a constructional or accommodation account anticipates the same possibility, and the source translations are passive-like despite the different hosts.

6. Discussion

6.1 Light-verb innovation, structural economy, and the finite-nonfinite asymmetry

The audited inventory supports an asymmetric bilingual predicate template: English-origin material contributes lexical and event meaning, whereas Pashto supplies overt tense, aspect, agreement, and clause organisation. The dominant kaw pattern instantiates the established matrix-language 'do' strategy; the Pashto-specific contribution is the category-neutral inventory and comparison of kaw with resultative/change-of-state keg/shw and copular hosts. The within-lemma use alternation further shows that host selection is not mechanically determined by the English lemma, although the archive cannot separate event construal, collocation, and discourse effects.

6.2 Comparison with alternative frameworks

Competing accounts remain viable. Matrix-language and borrowing typologies predict native-host accommodation; constructional approaches predict an entrenched [English-origin element + Pashto host] schema; lexical-selection and computational accounts predict effects of competition, semantic weight, and frequency (Purmohammad et al., 2022; Salig et al., 2024; Tsoukala et al., 2021; Wohlgemuth, 2009). The graded DEH adds explanatory value only if representational specification predicts residual host-choice or processing variation beyond these factors.

Persian-English provides the closest structural comparison, but Pashto offers an internal contrast among kaw, keg/shw, and copular integration. Translation-equivalence and support-verb diagnostics can refine that comparison. The DEH is therefore presented as a graded operationalisation of the 2014 low-competition proposal, distinguished by multiple-host predictions and explicit controls rather than by claiming the general economy insight as new.

Other frameworks remain genuine competitors. Distributed Morphology can represent roots and functional exponents separately (Halle & Marantz, 1993); Construction Grammar can encode a productive bilingual predicate template as a form-meaning pairing (Goldberg, 1995); borrowing accounts can derive light-verb accommodation; and bilingual production models can locate competition during lexical selection and grammatical encoding (Backus, 1996; Wohlgemuth, 2009; Purmohammad et al., 2022). Beyond the general low-competition prediction of Myers-Scotton and Jake (2014), the graded DEH contributes only if representational specification predicts variation beyond translation equivalence, support-verb correspondence, frequency, and constructional familiarity. Candidates with stronger English-origin argument-structural or inflectional requirements should show residual differences in host choice, disfluency, or latency relative to nominal-like candidates after the relevant controls are entered. Failure to observe those effects would remove the graded extension's explanatory advantage (Myers-Scotton & Jake, 2017; Treffers-Daller, 2025).

Borrowing status and cross-community variation restrict interpretation of the raw totals. Recurrence, listedness, phonological integration, dictionary status, Pashto equivalents, event domain, and discourse context could change the apparent distribution (Balam et al., 2024; Treffers-Daller, 2025). The kaw total may reflect constructional productivity, sampled topics, conventionalisation, or their interaction.

6.3 Broader implications for bilingual production and cognition

The pattern is compatible with lexical-conceptual contribution from English-origin material and finite realisation in Pashto, but English bare forms remain categorically ambiguous and some candidates may retain event- or argument-structural information. Controlled production should therefore test latency, disfluency, repair, and host choice while controlling frequency, borrowing, constructional familiarity, and adaptive language control.

6.4 Cross-linguistic predictions

Cross-linguistic comparisons are hypothesis-generating because published studies use different sampling, borrowing criteria, and denominators. Productive light verbs or verbalisers in Turkish, Persian, Tamil, Hindi/Urdu, and Brussels Dutch yield patterns qualitatively compatible with matrix-language finiteness. Chinese-English, Spanish-English, and Swahili-English illustrate how word order, limited obligatory verbal morphology, conjugational adaptation, or direct matrix-language affixation can produce different surface outcomes. The detailed comparison is provided in Supplementary Note S1, Supplementary Table S1; it does not constitute quantitative confirmation of graded processing cost.

6.5 From distribution to processing

The decisive test is not the occurrence of English-origin elements with inflected Pashto hosts, which several established accounts predict. It is whether independently coded representational specification explains host choice or processing after lemma frequency, borrowing, speaker, semantic domain, morphological availability, and constructional familiarity are controlled. A null residual effect would remove the graded DEH's proposed advantage.

6.6 Data transparency and reproducibility

Supplementary Data File S2 reports all 41 screened candidates, the six-item exclusion log, coding fields, available provenance, verification status, and a blank independent-recoding template. Supplementary Data File S1 preserves the complete retained interlinear Appendix A, including its morpheme, gloss, free-translation, and constituent-analysis tiers. High-confidence programme/reference mapping is available for 22 candidates; 19 lack exact programme-level mapping.

7. Limitations

The study has eight principal limitations. First, it concerns one language pair and a semi-formal public-broadcast register. Second, the retained archive documents Departmental Graduate Committee (DGC)

authorization, the clip inventory, selection process, Toolbox tiers, and reference-number system, but does not contain a separate ethics-reference number, original YouTube links, broadcast/access dates, total transcribed word count, or original media files. These are provenance and corpus-coverage limits: only English-containing sentences were retained for clause-based structural analysis, so the appendix was never a source-linked, exhaustive programme transcript. Token timestamps are not an analytical variable because the study asks how bilingual predicates function, not when switching occurs. Third, all 41 candidates occur in the portion of the archive designated as broadcast material, but only 22 have high-confidence programme/reference mappings; 19 lack exact programme-level mapping. Fourth, purposive selection of mixing-rich clips and English-containing sentences prevents prevalence estimates. Fifth, a boundary scan identified six additional NP + kaw sequences—study, analysis, table talk, discussion (two tokens), and calls—that resemble counted candidates at the coarse surface level but were excluded under the overt-nominal-NP rule; their forms and rationales are reported in Section 4.3.1 and the supplementary exclusion log. Sixth, independent coder sheets and adjudication records are unavailable, so reliability coefficients cannot be reported. Seventh, borrowing status cannot be resolved without recurrence, listedness, community, and phonological evidence. Eighth, host-opportunity denominators are unavailable. The claims are consequently limited to structural description of the broadcast-derived appendix inventory, not broadcast prevalence, processing causation, borrowing status, or community-wide grammatical preference.

8. Future Research

The most informative next study would combine an exhaustively sampled, source-linked corpus with controlled production. The corpus phase should record all eligible predicates, including Pashto-only controls, and model host/form choice against lemma frequency, argument structure, borrowing status, speaker, and programme. Complete programme transcripts and eligible-context denominators would support prevalence estimates; token-level temporal localisation could facilitate source checking but would not itself explain structural integration. A picture-description or structural-priming experiment could then compare English-origin nominal-like elements, bare verbal candidates, and manipulated finite conflicts while measuring latency, disfluency, repair, and host choice. The critical DEH result would be an effect of cross-language specification after frequency and constructional familiarity are controlled.

Cross-linguistic comparison should use commensurable token definitions and opportunity denominators rather than a qualitative supported/unsupported table. Pairs with productive light verbs, verbalising affixes, direct root affixation, and little obligatory verbal morphology offer a factorial design for separating morphological availability from election demand. Data, codebook, exclusions, and analysis scripts should be deposited where copyright and participant protection permit.

A further extension should test listedness and semantic-domain effects jointly: recurrence and association measures can operationalise conventionalisation following Treffers-Daller (2025), while balanced sampling across event domains can test the community-specific semantic pattern identified by Balam et al. (2024). Consistent with Rayo et al. (2024), participant language history and habitual

switching context should be reported with standardised measures so that structural and processing results are comparable across studies.

9. Conclusion

This study has audited 41 English-origin predicate candidates reconstructed from the complete retained Pashto-English Appendix A supplied as Supplementary Data File S1. Thirty-five constitute a conservative core set, five are category-ambiguous sensitivity cases, and one is excluded as a non-target motion-verb construction. Across all 40 hosted candidates, visible English tense or agreement morphology is absent and a Pashto light verb or copula carries the overt finite exponents. In the core set, the dominant configuration combines a bare English form with kaw 'do/make' (29/35), followed by keg/shw 'become' (5/35) and the copula (1/35). These are descriptive properties of the appendix-derived inventory and remain subject to audio and Toolbox verification.

The central contribution is an empirically delimited account of light-verb innovation and bilingual predicate creativity in Pashto-English. The general insight that EL nonfinite verbs and ML 'do' constructions may limit costly competition belongs to Myers-Scotton and Jake's (2014) production account. English-origin material expands the lexical-conceptual and event-semantic range of the Pashto VP, while Pashto hosts provide overt finite architecture and may contribute event and argument structure. The graded DEH formalises the resulting hypothesis that representational specification varies across candidates and should predict residual host-choice or processing effects. The current distribution is compatible with that extension but does not establish minimal production time, processing causation, host preference, or borrowing status. Those claims require controlled evidence after frequency, constructional familiarity, borrowing, provenance, reliability, and opportunity denominators are addressed (Balam et al., 2024; Treffers-Daller, 2025).

Abbreviations

1/2/3	first/second/third person
3SG	third person singular
ACC	accusative
AGR	agreement
BCV	bilingual compound verb
CL	clitic
COMP	complementizer
COP	copula

CS	code-switching
DAH	Differential Access Hypothesis
DEH	Differential Election Hypothesis
DEM	demonstrative
EL	embedded language
FUT	future
INF	infinitive
IPFV	imperfective
KP	Khyber Pakhtunkhwa
LVC	light verb construction
ML	matrix language
MLF	Matrix Language Frame (model)
NOM	nominative
OBL	oblique
PFV	perfective
PL	plural
PROG	progressive
PRS	present
PST	past
PTCP	participle
SBJV	subjunctive
SG	singular
SM	system morpheme
USP	Uniform Structure Principle

VBZ verbaliser

Declarations

Data Availability Statement

Three supplementary files are provided with this article: Supplementary Data File S1 (the complete retained interlinear Appendix A); Supplementary Data File S2 (the 41-candidate predicate audit, codebook, six-item screening-exclusion log, provenance fields, verification status, and blank independent-recoding template); and Supplementary Note S1 (the hypothesis-generating cross-linguistic comparison, including Supplementary Table S1). All analysed items derive from the public AVT Khyber broadcast portion of the archive; no privately elicited data are included. The dataset should therefore be treated as a broadcast-derived, clause-level appendix inventory, not a complete audiovisual corpus or a basis for prevalence estimates.

Use of Artificial Intelligence

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) to assist with language editing, clarity, organization, and the presentation of some parts of the analysis. The AI tool did not generate or alter the underlying research archive, linguistic examples, morpheme-level data, or analytical classifications. All examples, coding decisions, theoretical arguments, interpretations, citations, and conclusions were checked, revised, and approved by the authors, who accept full responsibility for the accuracy, originality, and integrity of the manuscript.

Acknowledgements

The author has no acknowledgements to declare.

Author Contributions

A.A.K. conceptualized the study, curated and audited the corpus data, developed the methodology, conducted the formal analysis, prepared the tables and figure, and drafted the manuscript. A.K. contributed to interpretation of the bilingual data and reviewed and revised the manuscript. Both authors approved the submitted version and accept accountability for the work.

Funding

This research received no external funding.

Competing Interests

The author declares no competing interests.

Ethical Approval and Accordance

Before data collection, the research plan was reviewed and approved through departmental academic oversight by the Departmental Graduate Committee (DGC), Department of Linguistics and Communications, School of Liberal Arts, University of Management and Technology, Lahore, Pakistan. This approval did not carry a separate institutional research-ethics/IRB reference number or formal exemption letter. The present article is a secondary linguistic analysis solely of publicly broadcast AVT Khyber programmes and involved no recruitment, private elicitation, intervention, or new interaction with individuals. No direct personal identifiers are included. All research was performed in accordance with the University of Management and Technology's applicable research policies and relevant Pakistani requirements for academic research, copyright, and attribution. The Declaration of Helsinki was not applicable because this was not medical research and no human participants were recruited for the present study.

Informed Consent

Informed consent was not required or obtained because the present study recruited no human participants, collected no private information, and analysed only material deliberately made public through AVT Khyber broadcasts. No new interaction with, intervention involving, or private elicitation from the broadcast speakers occurred.

Data Availability

Three supplementary files are provided with this article: Supplementary Data File S1 (the complete retained interlinear Appendix A); Supplementary Data File S2 (the 41-candidate predicate audit, codebook, six-item screening-exclusion log, provenance fields, verification status, and blank independent-recoding template); and Supplementary Note S1 (the hypothesis-generating cross-linguistic comparison, including Supplementary Table S1). They include retained morpheme segmentation, glosses, translations, host classifications, analytical status, available provenance, and verification status. All analysed items derive from the public AVT Khyber broadcast portion of the archive; no privately elicited data are included. High-confidence programme/reference mappings are available for 22 candidates, while 19 lack exact programme-level mapping. Original source links, broadcast recordings, Toolbox files, timestamps, total transcribed word count, and independent-coder records were not preserved and cannot be reconstructed. The dataset should therefore be treated as a broadcast-derived, clause-level appendix inventory, not a complete audiovisual corpus or a basis for prevalence estimates. Reuse remains subject to copyright, accurate attribution, and applicable institutional requirements.

References

1. Annamalai, E. (1989). The language factor in code-mixing. International Journal of the Sociology of Language, 75, 47–54.
2. Babrakzai, F. (1999). Topics in Pashto syntax [Doctoral dissertation]. University of Hawai□i at Mānoa.

3. Backus, A. (1996). Two in one: Bilingual speech of Turkish immigrants in the Netherlands. Tilburg University Press.
4. Balam, O., Pérez Leutza, L., Michalski, I., & Parafita Couto, M. C. (2024). A semantic analysis of bilingual compound verbs in two contact Spanish communities. Research in Corpus Linguistics, 12(1), 140–170. https://doi.org/10.32714/ricl.12.01.06
5. Bybee, J. (2010). Language, usage and cognition. Cambridge University Press.
6. Costa, A. (2005). Lexical access in bilingual production. In J. F. Kroll & A. M. B. De Groot (Eds.), Handbook of bilingualism: Psycholinguistic approaches (pp. 308–325). Oxford University Press.
7. De Bot, K. (1992). A bilingual production model: Levelt's 'speaking' model adapted. Applied Linguistics, 13(1), 1–24.
8. Francis, W. S. (2005). Bilingual semantic and conceptual representation. In J. F. Kroll & A. M. B. De Groot (Eds.), Handbook of bilingualism: Psycholinguistic approaches (pp. 251–267). Oxford University Press.
9. Garrett, M. F. (1975). The analysis of sentence production. In G. H. Bower (Ed.), The psychology of learning and motivation (Vol. 9, pp. 133–177). Academic Press.
10. Goldberg, A. E. (1995). Constructions: A construction grammar approach to argument structure. University of Chicago Press.
11. Green, D. W., & Abutalebi, J. (2013). Language control and the adaptive control hypothesis. Journal of Cognitive Psychology, 25(5), 515–530.
12. Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In K. Hale & S. J. Keyser (Eds.), The view from Building 20 (pp. 111–176). MIT Press.
13. Jiang, N. (2000). Lexical representation and development in a second language. Applied Linguistics, 21(1), 47–77.
14. Khalid, A., & Khan, A. A. (2022). Negotiation of bilingual identities: A case of young migrant Pashtuns in Lahore. Ampersand, 9, 100098.
15. Khan, A. A., & Khalid, A. (2018). Pashto–English code-switching: Testing the morphosyntactic constraints of the MLF model. Lingua, 201, 78–91.
16. Khan, A. A., Khalid, A., & Saleem, T. (2023). The evidence of embedded language islands: The case of Pashto–English code-switching. International Journal of Multilingualism. https://doi.org/10.1080/14790718.2023.2212907
17. Khan, A. A., & Muysken, P. (2014). Strategies for incorporating nouns and verbs in code-mixing: The case of Pashto–English bilingual speech. Lapurdum, 18, 97–137.
18. Kroll, J. F., & Tokowicz, N. (2005). Models of bilingual representation and processing: Looking back and to the future. In J. F. Kroll & A. M. B. De Groot (Eds.), Handbook of bilingualism: Psycholinguistic approaches (pp. 531–553). Oxford University Press.
19. Levelt, W. J. M. (1989). Speaking: From intention to articulation. MIT Press.

20. Lotfabbadi, L. N. (2002). Disagreement in agreement: A study of grammatical aspects of code-switching in Swedish–Persian bilingual speech [Doctoral dissertation]. Stockholm University.
21. Muysken, P. (2000). Bilingual speech: A typology of code-mixing. Cambridge University Press.
22. Muysken, P. (2016). From Colombo to Athens: Areal and universalist perspectives on bilingual compound verbs. Languages, 1(1), Article 2. https://doi.org/10.3390/languages1010002
23. Myers-Scotton, C. (1993). Duelling languages: Grammatical structure in code-switching. Oxford University Press.
24. Myers-Scotton, C. (2002). Contact linguistics: Bilingual encounters and grammatical outcomes. Oxford University Press.
25. Myers-Scotton, C. (2006). Multiple voices: An introduction to bilingualism. Blackwell.
26. Myers-Scotton, C., & Jake, J. L. (2000). Four types of morpheme: Evidence from aphasia, code-switching and second-language acquisition. Linguistics, 38(6), 1053–1100.
27. Myers-Scotton, C., & Jake, J. L. (2009). A universal model of code-switching and bilingual language processing and production. In B. E. Bullock & A. J. Toribio (Eds.), The Cambridge handbook of linguistic code-switching (pp. 336–357). Cambridge University Press.
28. Myers-Scotton, C., & Jake, J. L. (2014). Nonfinite verbs and negotiating bilingualism in codeswitching: Implications for a language production model. Bilingualism: Language and Cognition, 17(3), 511–525. https://doi.org/10.1017/S1366728913000758
29. Myers-Scotton, C., & Jake, J. L. (2017). Revisiting the 4-M model: Code-switching and morpheme election at the abstract level. International Journal of Bilingualism, 21(3), 340–366.
30. Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en español: Toward a typology of code-switching. Linguistics, 18(7–8), 581–618.
31. Purmohammad, M., Vorwerg, C., & Abutalebi, J. (2022). The processing of bilingual (switched) compound verbs: Competition of words from different categories for lexical selection. Bilingualism: Language and Cognition, 25(5), 755–767. https://doi.org/10.1017/S1366728921001103
32. Rayo, W., Barrita, A. M., Cabrera Martinez, L., & Carbajal, I. (2024). The complexity in bilingual code-switching research: A systematic review. Languages, 9(6), 217. https://doi.org/10.3390/languages9060217
33. Sankoff, D., Poplack, S., & Vanniarajan, S. (1990). The case of the nonce loan in Tamil. Language Variation and Change, 2(1), 71–101.
34. Salig, L. K., Valdes Kroff, J. R., Slevc, L. R., & Novick, J. M. (2024). Linking frequency to bilingual switch costs during real-time sentence comprehension. Bilingualism: Language and Cognition, 27(1), 25–40. https://doi.org/10.1017/S1366728923000366
35. Siegel, J. (1987). Language contact in a plantation environment: A sociolinguistic history of Fiji. Cambridge University Press.
36. Treffers-Daller, J. (1994). Mixing two languages: French–Dutch contact in a comparative perspective. Mouton de Gruyter.

37. Treffers-Daller, J. (2025). The Simple View of borrowing and code-switching. International Journal of Bilingualism, 29(2), 347–370. https://doi.org/10.1177/13670069231168535
38. Tsoukala, C., Frank, S. L., Van den Bosch, A., Valdes Kroff, J., & Broersma, M. (2021). Modeling the auxiliary phrase asymmetry in code-switched Spanish-English. Bilingualism: Language and Cognition, 24(2), 271–280. https://doi.org/10.1017/S1366728920000449
39. Wei, L. (2001). Lemma congruence checking between languages as an organizing principle in intrasentential code-switching. International Journal of Bilingualism, 5(2), 153–173.
40. Wei, L. (2002). The bilingual mental lexicon and speech production process. Brain and Language, 81(1–3), 691–707.
41. Wohlgemuth, J. (2009). A typology of verbal borrowings. Mouton de Gruyter.

Figures

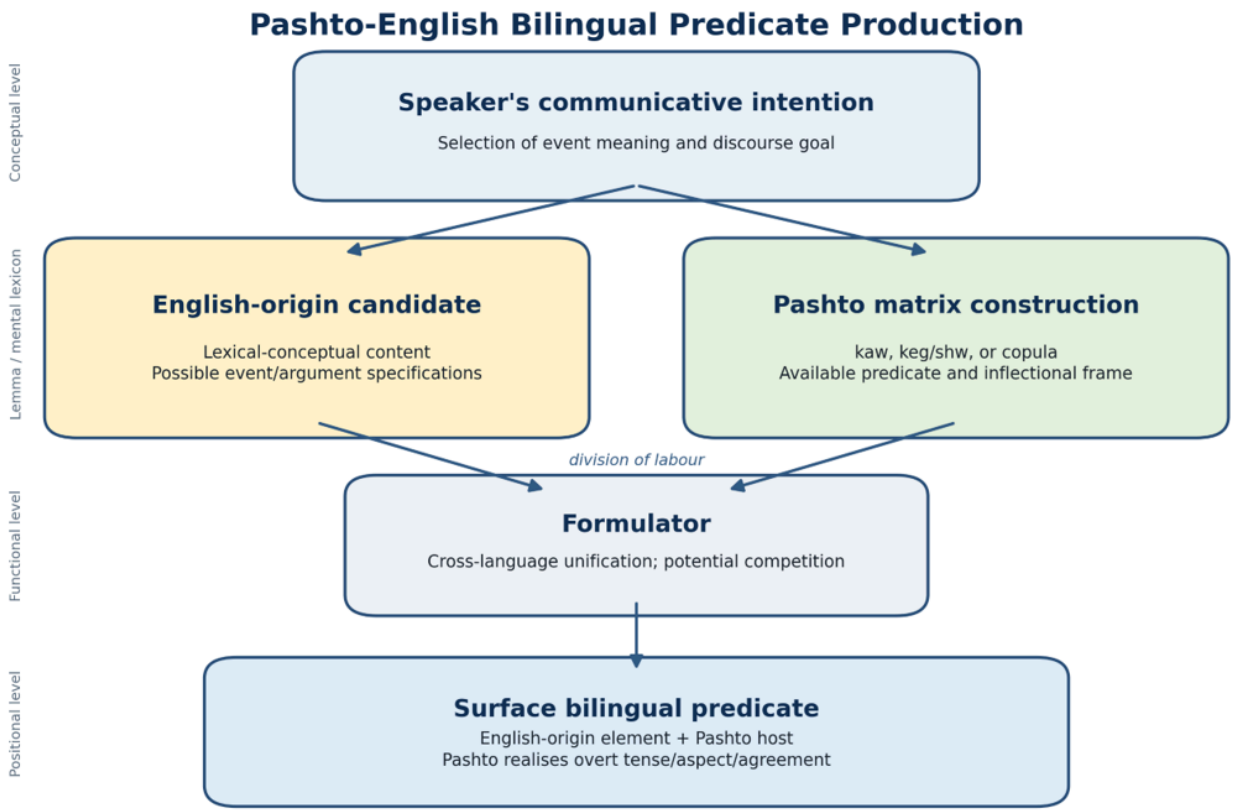


Figure 1

Pashto-English bilingual predicate-production pathway. Note. Author-developed application of the production architecture proposed by Myers-Scotton and Jake (2014, p. 516); it is not a reproduction of their figure.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryDataS1CompleteRetainedInterlinearAppendixA.docx
- SupplementaryNoteS1CrossLinguisticDEHComparison.docx
- SupplementaryDataS241ItemPredicateAuditandCodebook.xlsx
- AdditionalInformation.docx