

DATA-DRIVEN LEARNING AND ACADEMIC COLLOCATIONAL COMPETENCE IN EFL: A CORPUS-INFORMED INSTRUCTIONAL FRAMEWORK FOR B2-C1 LEARNERS

Ziyayeva Muxarramxon Nasimxo‘ja qizi
umarkhujayevamuharamkhon@gmail.com

Abstract: Collocational competence is now recognised as a principal threshold separating formally accurate from genuinely idiomatic academic English, yet it remains the least systematically taught component of the EFL curriculum. The article addresses the persistent collocation deficit of advanced learners by designing and theoretically validating a four-phase corpus-informed instructional framework (Notice - Explore - Formulate - Apply, NEFA) for the B2-C1 range. The study combines documentary analysis of reference corpora, academic word and collocation lists and CEFR Companion Volume descriptors with a contrastive congruence analysis of first-language-induced miscollocation.

Keywords: data-driven learning; collocational competence; corpus linguistics; academic writing; English for Academic Purposes; L1 transfer; noticing; concordancing; CEFR

Introduction

Forty years of phraseological research have converged on a single, methodologically inconvenient fact: fluent language is far less freely generated than grammars imply. J. Sinclair’s distinction between the open-choice principle and the idiom principle established that a large proportion of running text is produced from semi-preconstructed multi-word units rather than assembled slot by slot [1]. Subsequent corpus work has consolidated the claim that mastery of conventionalised co-occurrence - collocation - is not an ornamental refinement added after grammar and vocabulary, but a constitutive component of proficiency [9; 22].

The pedagogical consequence is uncomfortable. Learner-corpus studies show that collocational accuracy does not improve in step with general proficiency: N. Nesselhauf demonstrated that advanced learners continue to produce deviant verb-noun combinations at a rate disproportionate to their grammatical control, and that the single strongest predictor of error is non-congruence between the L1 and L2 combination [4]. B. Laufer and T. Waldman replicated the finding across proficiency bands and added a longitudinal dimension: the number of collocations produced grows with proficiency, while the proportion of erroneous ones does not fall correspondingly, indicating that exposure alone does not restructure the learner’s phrasicon [5]. S. Granger’s analysis of prefabricated patterns in advanced EFL writing points in the

same direction, showing simultaneous overuse of a narrow set of safe amplifiers and underuse of the register-appropriate repertoire [13]. Data-driven learning (DDL), proposed by T. Johns, offers a principled response: instead of receiving usage rules from the teacher, learners interrogate authentic concordance data and induce the patterns themselves [2]. The empirical record is favourable. A. Boulton and T. Cobb's meta-analysis of sixty-four studies reports a medium-to-large aggregated effect for DDL over comparison conditions, with the largest gains precisely in the area of lexicogrammatical patterning and collocation [3]. A. Boulton has further shown that the effect does not depend on sophisticated software: printed, teacher-prepared concordances produce comparable gains, which makes DDL viable in technologically constrained classrooms [19]. Two gaps motivate the present study. First, the DDL literature is dominated by Western European and East Asian settings; little has been published on adapting the approach to Turkic-L1 contexts, where the congruence profile of verb-noun collocations differs substantially from that of the Indo-European languages on which the standard error taxonomies were built. Second, the instructional landscape has changed: generative language models now answer usage questions instantly and fluently, and learners increasingly consult them instead of corpora. Since such models produce plausible rather than attested output and provide no frequency evidence, the relationship between corpus consultation and model consultation requires explicit methodological regulation rather than tacit avoidance.

Materials and Methods

The study is a theoretical-constructive (design-based) inquiry. Its product is not a statistical generalisation but an artefact - an instructional framework - derived from theory, constrained by documentary evidence and specified in sufficient detail to be implemented and subsequently tested. Accordingly, the Results section reports the designed artefact and its evidential warrants; no learner performance data were collected, and no effectiveness claims are advanced in advance of the trial outlined in subsection 4.4. Four classes of documentary material were used. (i) Reference corpora and query interfaces: the Corpus of Contemporary American English and its academic subcorpus [17], the British National Corpus, the Sketch Engine platform with its word-sketch and thesaurus functions [18], the pedagogically simplified SKELL interface, and the offline concordancer AntConc [20]. (ii) Lexical inventories: the Academic Word List [6] and the Academic Collocation List, derived from the written curricular component of the Pearson International Corpus of Academic English [7]. (iii) Descriptor systems: the CEFR Companion Volume, in particular the scales for vocabulary range and control and for written production at B2, B2+ and C1 [16]. (iv) Theoretical and empirical sources on phraseology in second language acquisition [1; 4; 5; 8; 9; 13; 22].

Four procedures were applied in sequence. *Documentary analysis* of the descriptor scales established the competence targets against which the framework outputs are aligned. *Contrastive congruence analysis* compared English academic verb-noun and adjective-noun combinations with their closest Uzbek functional counterparts, classifying each pair as congruent, partially congruent or non-congruent, following the procedure established for learner-corpus error analysis [4]. *Taxonomic construction* grouped the resulting non-congruence patterns into error types and tested each type for diagnostic distinctness. *Theoretical alignment* mapped the phases of the framework onto the noticing hypothesis [10], the output hypothesis [11] and task-based sequencing principles [12]. The English collocations cited below are drawn from the academic lists and reference corpora named in 2.2. The deviant forms marked with an asterisk are *typologically representative constructions* illustrating documented transfer mechanisms; they are presented as analytic exemplars of a pattern, not as frequency counts from a learner corpus, and no numerical claims are attached to them. Establishing their distribution in Uzbek learner writing is an explicit task of the empirical stage described in 4.4.

Results and Discussion

Analysis of the non-congruence patterns yielded six diagnostically distinct types (Table 1). The organising variable is the relation between the L1 combinatorial template and the L2 norm: the greater the surface similarity between the two languages, the higher the probability that the learner will transfer the L1 template unmodified - a paradox already noted in the learner-corpus literature, where congruence facilitates and near-congruence misleads [4; 5].

Table 1.

Congruence-based typology of academic miscollocations (English, B2-C1)

Type	Mechanism	L1 template (Uzbek)	Deviant form	Academic norm
I. Delexical verb substitution	The L1 uses one general-purpose verb where English distributes several delexical verbs	<i>tadqiqot qilmoq</i> 'to do research'	make a research	conduct / carry out research
II. Literal verb calque	The L1 verb is translated by its primary dictionary equivalent irrespective of collocational range	<i>maqsad qo'ymoq</i> 'to put a goal'	put a goal	set / pursue an objective

III. Converse-role reversal	The L1 encodes the participant roles of the event in the opposite direction	<i>imtihon topshirmoq</i> 'to hand in an exam'	give an exam (of a student)	take / sit an examination
IV. Intensifier mismatch	A semantically 'strong' L1 adjective is mapped onto the wrong English amplifier	<i>kuchli yomg'ir</i> 'strong rain'; <i>katta e'tibor</i> 'big attention'	strong rain; big attention	heavy rain; close / considerable attention
V. Register-level mismatch	A formally correct combination belongs to a lower register than the academic genre requires	- (intralingual)	the article talks about...; a lot of data	the article examines / addresses...; a substantial body of data
VI. Support-noun omission / addition	Nominalised academic patterns require support elements absent in the L1 template	<i>xulosa chiqarmoq</i> 'to take out a conclusion'	take out a conclusion	draw / reach a conclusion

Three properties make the typology usable as a diagnostic instrument. It is exhaustive at the level of the verb-noun and adjective-noun axis that dominates academic phraseology; it is contrastively motivated, so that each type predicts a class of errors before it occurs rather than merely labelling it afterwards; and each type maps onto a distinct corpus query. Type I and Type VI are diagnosed by a node-noun query with verb collocates ranked by mutual information; Type IV by an adjective-collocate sketch of the node noun; Type V by a comparison of frequencies across the academic and spoken subcorpora - a contrast that makes register gradience visible to learners who otherwise perceive lexis as binary correct/incorrect.

The framework organises corpus consultation into four phases, each defined by a theoretical warrant, a tool, a learner action, an observable product and an assessment criterion (Table 2). Its central design decision is that no phase terminates in recognition: every phase yields an artefact that can be inspected, and the cycle closes in obligatory pushed output.

Table 2.

The NEFA framework: phases, warrants, tools and products

Phase	Theoretical warrant	Tool / input	Learner action	Observable product
-------	---------------------	--------------	----------------	--------------------

N - Notice	Noticing hypothesis: intake requires conscious registration of form [10]	Teacher-prepared KWIC extract (12-15 lines) of a high-frequency academic node noun	Underline the verb slot in each line; group recurrent verbs	Annotated concordance sheet
E - Explore	Inductive rule discovery; learner-as-researcher [2; 19]	COCA academic subcorpus / SKELL word sketch [17; 18]	Compare collocate profiles of the node across academic and spoken registers	Comparative collocate table with a one-sentence generalisation
F - Formulate	Restructuring of the interlanguage rule; explicit-implicit interface	Class discussion; Academic Collocation List as a check [7]	State the pattern as a usage rule and record the L1-L2 congruence status	Personal collocation notebook entry (pattern + counter-example)
A - Apply	Output hypothesis: production forces syntactic processing [11]; task-based cycle [12]	Own draft paragraph; corpus for verification	Rewrite a paragraph of one's own text using the target patterns; verify each in the corpus	Revised paragraph plus a corpus-verified revision log

The phases are cumulative rather than optional. Removing

Notice reduces the sequence to unguided browsing, which the DDL literature identifies as the principal cause of learner disengagement and cognitive overload [15]. Removing *Formulate* leaves the induced pattern implicit and therefore unavailable for monitoring. Removing *Apply* converts the lesson into a receptive exercise, reproducing precisely the recognition-production gap that the intervention is designed to close.

Generative language models cannot be excluded from the contemporary classroom, and a prohibition would merely drive their use out of sight. The protocol therefore assigns them a bounded, secondary role defined by three rules. Rule 1 (primacy of attestation): a collocation enters the learner's notebook only after it has been observed in corpus data; model output may suggest candidates but cannot license them. Rule 2 (frequency evidence): any claim about relative typicality - which of two variants is preferred in academic prose - must be settled by a frequency comparison,

since models return no distributional evidence and their fluency is uncorrelated with corpus frequency. Rule 3 (register localisation): register judgements are made by cross-subcorpus comparison, not by asking a model whether a phrase ‘sounds formal’. Pedagogically, the protocol converts a threat into a productive contrast: the model becomes a hypothesis generator whose output the learner is required to falsify or confirm against evidence. This preserves the epistemological core of DDL - the learner as investigator of attested language - while acknowledging the tool that learners actually use. The typology in Table 1 makes an empirical prediction that distinguishes it from inventories organised by part of speech: error probability is not a function of item difficulty but of the *distance* between the L1 and L2 templates, and it is highest in the zone of near-congruence. **Make a research* is produced not because the learner lacks the verb *conduct*, but because the Uzbek template *tadqiqot qilmoq* offers a superficially adequate mapping that the learner has no reason to distrust. This is consistent with Nesselhauf’s finding that congruent combinations are produced accurately, non-congruent ones are often avoided, and partially congruent ones generate the bulk of errors [4]. It follows that instruction should be targeted at the near-congruent zone rather than distributed evenly across the lexicon - a resource-allocation argument of direct practical value in contexts with limited contact hours.

The recurrent criticism of DDL is that learners notice patterns without gaining productive control of them. The NEFA design addresses this at the level of task architecture rather than exhortation. Phase N satisfies Schmidt’s condition that intake presupposes conscious registration [10]; the prediction-attestation mismatch is the mechanism by which registration is forced rather than hoped for. Phase A satisfies Swain’s condition that production pushes the learner from semantic to syntactic processing [11], and the requirement that the rewritten text be the learner’s *own* is not incidental: it locates the new pattern in a communicative context the learner is already committed to, which is the condition under which restructuring rather than memorisation occurs [12]. Direct corpus access is not uniformly available. Boulton’s evidence that paper-based concordances yield gains comparable to software-mediated ones [19] is therefore central to the framework’s transferability: phases N and F require no connectivity, and phases E and A can be run on teacher-prepared printouts at the cost of some learner autonomy. What cannot be economised is teacher preparation. Query design is a professional skill, and the framework presupposes a teacher able to select nodes, filter irrelevant lines and anticipate misleading hits - an implication for in-service teacher education at least as significant as the classroom procedure itself [14].

The study delivers a designed artefact, not an effectiveness claim; the framework has not yet been trialled, and no numerical results are reported or implied here. The transfer patterns of Table 1 are typologically motivated exemplars whose actual

distribution in Uzbek learner writing remains to be established. Validation should therefore proceed in two stages. Stage one: compilation of a learner corpus of academic essays written by Uzbek-L1 students at B2-C1, annotated for miscollocation using the six-type scheme, with inter-rater agreement reported. Stage two: a quasi-experimental pre-test / post-test / delayed-post-test comparison of NEFA instruction against a matched explicit-teaching condition over a full semester, using the productive instruments of 3.5, with effect sizes and confidence intervals reported and the analysis plan registered in advance. Only after stage two can claims of comparative effectiveness be advanced.

Conclusion: The article has argued that the collocational deficit of advanced EFL learners is a structural rather than an incidental problem, and that it is tractable through a disciplined form of corpus consultation. Three results were presented. First, a six-type congruence-based typology of academic miscollocations that predicts error classes from L1-L2 template distance and maps each class onto a specific corpus query. Second, the NEFA framework, which binds each phase of consultation to a theoretical warrant, a tool, an observable learner product and an assessment criterion, thereby closing the recognition-production gap that has limited earlier DDL implementations. Third, a corpus-first verification protocol that admits generative language models as hypothesis generators while reserving evidential authority for attested, frequency-annotated data.

References

1. Sinclair J. Corpus, Concordance, Collocation. - Oxford: Oxford University Press, 1991. - 179 p.
2. Johns T. Should you be persuaded: Two samples of data-driven learning materials // ELR Journal. - Birmingham: University of Birmingham, 1991. - Vol. 4. - P. 1-16.
3. Boulton A., Cobb T. Corpus use in language learning: A meta-analysis // Language Learning. - 2017. - Vol. 67, No. 2. - P. 348-393.
4. Nesselhauf N. The use of collocations by advanced learners of English and some implications for teaching // Applied Linguistics. - 2003. - Vol. 24, No. 2. - P. 223-242.
5. Laufer B., Waldman T. Verb-noun collocations in second language writing: A corpus analysis of learners' English // Language Learning. - 2011. - Vol. 61, No. 2. - P. 647-672.
6. Coxhead A. A new academic word list // TESOL Quarterly. - 2000. - Vol. 34, No. 2. - P. 213-238.

7. Ackermann K., Chen Y.-H. Developing the Academic Collocation List (ACL): A corpus-driven and expert-judged approach // Journal of English for Academic Purposes. - 2013. - Vol. 12, No. 4. - P. 235-247.
8. Lewis M. The Lexical Approach: The State of ELT and a Way Forward. - Hove: Language Teaching Publications, 1993. - 200 p.
9. Nation I.S.P. Learning Vocabulary in Another Language. - Cambridge: Cambridge University Press, 2001. - 477 p.
10. Schmidt R. The role of consciousness in second language learning // Applied Linguistics. - 1990. - Vol. 11, No. 2. - P. 129-158.