

MEDLEX25: DIAGNOSING THE RECOGNITION-PRODUCTION GAP IN MEDICAL ENGLISH VOCABULARY

Evgeni Stanchev

Medical University of Varna, Bulgaria

Abstract



This study introduces MedLex-25, a compact diagnostic instrument designed to quantify recognition-production asymmetries in medical English vocabulary. Second-year medical students at Medical University Varna, Bulgaria completed nine parallel test forms (50 targets per form; 225 unique lexical targets sampled from a coursebook corpus) under supervised conditions; usable data from 186 students were analysed, with a total of 9300 questions answered. Each form paired a four-option receptive item with a sentence-production task for the same targets; production responses were binary-scored (correct/incorrect) with tolerance for non-meaning-preventing surface errors. Results suggest a systematic recognition-production gap across five lexical categories (nouns, verbs, adjectives, adverbs, two-word multiword items): receptive accuracy consistently exceeded productive accuracy, and no item achieved perfect production. Counts of perfectly recognized items were: nouns 33/45, adjectives 27/45, verbs 23/45, adverbs 25/45, and multiword items 28/45. Verbs, adjectives and—most strikingly—adverbs exhibited the largest productive weaknesses (lowest production for erroneously = 0.071). Item-level diagnostics identify concrete teaching targets. The MedLex-25 test can guide focused instruction and recommend follow-up studies that expand item coverage, refine production scoring, and investigate etymology and cross-disciplinary exposure as mediating factors.

Keywords: Medical English, ESP vocabulary, recognition-production gap, error analysis, corpus-based item selection

Article history:

Received: 18 December 2025

Reviewed: 12 January 2026

Accepted: 10 March 2026

Published: 22 June 2026

Copyright © 2026 Evgeni Stanchev



This is an Open Access article published and distributed under the terms of the [CC BY 4.0 International License](https://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Citation: Stanchev, E. (2026). MedLex25: Diagnosing the recognition-production gap in medical English vocabulary. *English Studies at NBU*, 12(1), 113-132. <https://doi.org/10.33919/esnbu.26.1.8>

Evgeni Stanchev is a lecturer at the Department of Foreign Languages, Communication and Sports at Medical University – Varna. He holds an MA in English Philology (Linguistics and Translation) and is currently PhD candidate with a research focus on “The role of AI in teaching ESP for medicine and healthcare”. His main interests include linguistics, technology and their practical application within a modern-day classroom.

E-mail: evgeni.stanchev@mu-varna.bg

 <https://orcid.org/0009-0001-1108-1067>

Effective communication in clinical settings depends crucially on domain-specific vocabulary: medical students must both recognise technical terms in texts and produce them accurately in writing and speaking. Research consistently indicates that receptive knowledge (recognition) tends to outstrip productive knowledge (use) in L2 vocabulary learning, a pattern that has practical implications for teaching in English for Specific Purposes (ESP) courses for medicine. Recent empirical work confirms substantial receptive–productive asymmetries across contexts, indicating that learners often understand more terms than they can reliably deploy in context (Malmström et al., 2025).

For medical ESP this asymmetry is particularly consequential: imprecise or incorrect lexical production can distort clinical meaning and potentially endanger patient care. Consequently, there is a pressing need for short, reliable diagnostic instruments that (a) sample domain-relevant vocabulary ecologically, and (b) distinguish recognition from production at the item level so instructors can prioritise teaching targets. Corpus-based selection of items is widely accepted as the most defensible way to ensure domain relevance, but corpus analysis alone does not reveal whether students can productively use the items in discourse. Recent studies in collocation and lexical acquisition underline that learning outcomes depend on both exposure and task design, and that productive mastery requires pedagogical practice beyond mere recognition (Pu et al., 2024).

This study introduces MedLex-25, a compact diagnostic instrument built from coursebook-derived corpus data and designed to measure recognition and production of medical English across five lexical categories (nouns, verbs, adjectives, adverbs, and two-word multiword items). Unlike corpus projects that focus primarily on the compilation, annotation, and analysis of medical texts, MedLex-25 uses corpus evidence as the item-selection stage and then operationalises those items through paired receptive (multiple-choice) and productive (sentence-production) tasks completed by learners. In this way, the instrument adds a pedagogical and measurement layer that moves beyond corpus description toward classroom-based lexical diagnosis. While MedLex-25 is informed by corpus methodology and shares conceptual affinities with earlier MEDLEX projects in its domain focus and principled lexical sampling, it is not a descendant of those corpus

initiatives and differs fundamentally in aim and procedure, shifting the emphasis from corpus construction to learner performance and diagnostic use (Kokkinakis, 2006).

The present study does not aim to provide a complete psychometric validation of MedLex-25. Rather, it reports the initial empirical piloting and exploratory validation of an early-stage diagnostic instrument, developed by the author as part of the broader research and pedagogical work in medical English for specific purposes. MedLex-25 was conceived as a formative, classroom-diagnostic tool intended to support instructional decision-making, rather than as a standardized or high-stakes assessment instrument. Accordingly, the purpose of this study is to examine whether the instrument can function adequately in authentic classroom conditions, elicit stable and interpretable recognition–production asymmetries, and differentiate lexical categories in theoretically meaningful ways. The findings are therefore presented as initial evidence of diagnostic adequacy in a pilot context, intended to inform further refinement and future large-scale validation, rather than as claims of definitive instrument validation.

Research questions

The empirical work is guided by two primary research questions that connect diagnostic measurement to pedagogical decision-making:

1. Which lexical category—nouns, verbs, adjectives, adverbs, or two-word multiword items—exhibits the largest receptive–productive gap among medical students assessed with MedLex-25?
2. Which factors account for observed receptive–productive gaps across lexical categories (e.g., item frequency in the coursebook corpus, collocational constraints, part-of-speech-specific processing demands, and task affordances)?

Expected contributions are twofold: (i) identification of lexical categories most susceptible to productive breakdowns in medical English; and (ii) specification of candidate explanatory mechanisms that can inform targeted pedagogical interventions and further refinement of the diagnostic instrument. Subsequent sections describe MedLex-25 development and validation, report classroom findings, and discuss implications for ESP assessment and instruction in medical contexts.

Literature Review

Recognition vs. production: empirical patterns and educational stakes

Receptive-productive distribution patterns are similar across contexts. This author's earlier empirical investigations are consistent with the idea that learners perform better on receptive tasks (multiple-choice, recognition) than on production tasks (gap-fill, sentence generation), and that the size of the gap is sensitive to a variety of factors. For example, large-sample studies of vocabulary in instructional contexts show that even experienced learners and teachers have larger receptive than productive inventories, a reality that has direct consequences for ESP syllabi which often emphasise reading comprehension at the expense of controlled production practice (Xulong & Sukying, 2021).

Collocational knowledge (the habitual pairing of words) has been shown to be particularly sensitive to the receptive–productive divide. In a longitudinal study tracking L2 German learners over three years, learners gradually shifted some collocations from passive (receptive) recognition to active (productive) use, depending on factors such as item frequency, congruency, and prior productive vocabulary knowledge (Boone et al., 2023). Cross-sectional data from Thai university learners similarly reveal significantly higher receptive than productive collocation knowledge across proficiency levels (Sonbul et al., 2023). Furthermore, an intervention using spaced retrieval practice of verb–noun collocations resulted in much stronger long-term productive collocation retention than massed schedules, indicating that repeated focused retrieval supports productive mastery (Yamagata et al., 2023). Educational research also documents generally low collocational competence among L2 learners, with productive collocation knowledge frequently lagging behind receptive knowledge — a pattern that underscores the utility of diagnostic instruments combining both receptive and productive measures (Begagić, 2025).

Corpus-based item selection in ESP diagnostics

Corpus analysis is the methodological backbone of domain-specific item selection: medical corpora allow test designers to identify high-value lexical items, typical multiword expressions, and the collocations that characterise discipline-specific discourse. The MEDLEX family of projects (notably Kokkinakis's Swedish MEDLEX corpus

work) illustrates how carefully compiled and annotated medical corpora can support lexicon building and linguistic analysis for biomedical language processing (Kokkinakis, 2006). More recent corpus-driven ESP research shows the pedagogical potential of such resources: multimodal medical-English corpora have been integrated into instruction with measurable proficiency gains (Kang & Niu, 2024), and corpus-derived pedagogic word lists have been operationalised for materials development and learner support in specialised domains (Fraser et al., 2025). Experimental applications of corpus-derived vocabulary sets in ESP classrooms similarly suggests that corpus linguistics methods can be extended beyond descriptive analysis into empirically validated teaching and assessment tools (Đurović & Bauk, 2022). Furthermore, an intervention using spaced retrieval practice of verb–noun collocations resulted in much stronger long-term productive collocation retention than massed schedules, indicating that repeated focused retrieval supports productive mastery (Yamagata et al., 2023). MedLex-25 intentionally bridges these stages by using corpus frequency and distribution to select targets and then administering paired receptive/productive tasks to students. In doing so, MedLex-25 extends the corpus approach with an empirical classroom layer that measures actual learner performance on corpus-derived items.

Corpus-driven approaches have played a central role in establishing empirically grounded lexical inventories for medical English, ensuring that target items are selected on the basis of authentic disciplinary usage rather than pedagogical intuition. Research grounded in specialised medical corpora has demonstrated how systematic lexical extraction and categorisation can illuminate recurrent lexical patterns, including verb choice and distribution in medical academic discourse (Doykova, 2017). Such work provides a descriptive foundation for ESP materials development and assessment by identifying high-value lexical items and phraseological tendencies characteristic of medical texts. At the same time, corpus-based analyses are primarily concerned with the distribution and functions of lexical items within expert discourse and do not directly address how effectively these items are recognised or produced by learners in instructional contexts. Building on this corpus-informed tradition, MedLex-25 retains corpus-based item selection to ensure domain relevance but extends it through paired receptive and productive testing, thereby operationalising corpus-derived vocabulary as

a diagnostic instrument which may help identify category-specific and item-level production bottlenecks in medical ESP learners.

Category-specific vulnerabilities: verbs, collocations and adverbials

Empirical research on vocabulary acquisition repeatedly identifies verbs, collocations, and adverbials as categories that pose particular production challenges. Verbs convey argument structure/- and subcategorisation possibilities that learners must master to produce functionally accurate clauses (Nakamura et al., 2022). Collocations demand associative strength and interlexical knowledge that often fails to generalise from recognition to use; large-sample and experimental studies document sizeable receptive–productive gaps for collocations and emphasise measurement and item-format effects (Tran & Waluyo 2021). Adverbials, because of their syntactic mobility and subtle semantic/pragmatic effects, frequently trigger placement and selection errors in learner output (Miller & Larsson 2024). Intervention research indicates that targeted task designs — for example, repetition with variation, spaced/interleaved retrieval practice, receptive–productive integration tasks, and explicit collocation training — are effective in strengthening productive knowledge and reducing receptive–productive gaps (Rogers, 2025). Instruction of this kind remains unevenly implemented in medical ESP curricula, limiting transfer of corpus-derived targets into controlled productive competence (Nguyen, 2025). These findings motivated testing of five discrete lexical categories and the reporting of item-level error profiles to guide concrete pedagogical responses.

Diagnostic instruments and recent methodological advances

Recent methodological work emphasises two complementary desiderata for vocabulary diagnostics: (1) ecological validity — items should reflect authentic domain usage, typically achieved through corpus-informed selection, and (2) diagnostic resolution — paired receptive/productive items and item-level error analysis, which can identify specific instructional targets, such as argument-structure issues or collocational misuse (Durrant, 2014). Technological affordances (e.g., online administration platforms, automated export of learner responses) further enable scalable, low-stakes testing, while emerging analytical practices—such as reporting counts of perfectly recognised items and flagging systematically low-performing targets—provide actionable insights for

instructors (Edmonds et al., 2022). MedLex-25 adopts both principles: corpus-informed item choice enhances domain relevance, and a two-stage receptive–production format captures the production bottleneck characteristic of specialised vocabulary learning (Mohammadi et al., 2024).

MedLex-25 is therefore best understood as a hybrid: it is grounded in corpus evidence (to ensure that tested items are relevant and representative of medical discourse) but extends corpus analysis by operationalising those items in paired receptive and productive tasks and then analysing actual learner output. This second layer — the empirical measurement of production against recognition — is crucial because corpus frequency alone does not predict producibility in L2 learners (Webb, 2008; Lee, 2025); only by testing students on corpus-derived items under controlled conditions can one identify which items are merely familiar and which are functionally available for use. In short, MedLex-25 combines corpus constitution with pedagogical diagnosis, moving from description toward prescriptive classroom guidance (Fang et al., 2021; Rasikawati, 2019).

Method

Study design and aims

This study is a follow-up to earlier work, which employed paired receptive and productive tasks to demonstrate a systematic gap between medical students' ability to recognize specialized vocabulary and their ability to produce it accurately in context. While that study provides initial evidence of the existence of a recognition–production mismatch in Medical ESP, it did not differentiate difficulty across lexical categories. The present study extends this line of inquiry by systematically examining which parts of speech in medical lexis—nouns, verbs, adjectives, adverbs, and two-word multiword items — pose the greatest challenges for second-year medical students in both recognition and production. Using the MedLex diagnostic instrument, constructed from coursebook-based corpus data, the study combines receptive (multiple-choice) and productive (sentence-production) measures to quantify the recognition–production gap by lexical category.

Participants

Participants were 186 second-year medical students at Medical University Varna, Bulgaria. All participants were native speakers of Bulgarian; a subset attended language-specialist secondary schools and reported CEFR levels in the B1–B2 range. Testing took place during regular class sessions; students completed the instrument under supervised conditions.

Materials and item selection

The source material was the approved and currently used student coursebook for the Medical English course, *Professional Communication in English: Medicine*, authored by Assoc. Prof. Ilina Doykova (2022; ISBN 978-619-221-398-5). For the purposes of this study, the coursebook was compiled into a corpus using Sketch Engine. From this corpus, target items were extracted across five lexical categories: nouns, verbs, adjectives, adverbs, and two-word noun *n*-grams.

For each lexical category within each test form, five target items were selected based on corpus frequency: one highest-frequency item, one lowest-frequency item, and three mid-frequency items. Nine parallel test forms were constructed, each containing a unique set of 25 target items (5 items × 5 categories). The use of nine non-overlapping forms yielded a total of 225 unique lexical targets, thereby increasing lexical coverage and enhancing the robustness of the instrument.

Instrument development rationale

The MedLex-25 was developed in response to a practical pedagogical need in medical ESP instruction: instructors require a classroom-feasible way to identify which corpus-relevant lexical items students can recognise but cannot reliably produce in context, so that instruction can be targeted toward the most consequential productive gaps. The instrument therefore prioritises formative, item-level diagnostics that can be administered within regular teaching time and used to guide instructional focus rather than summative evaluation.

Design choices were informed by prior classroom observation and by earlier paired receptive–productive work in medical ESP (Stanchev, 2025), which suggested that diagnostic instruments are most informative when they avoid extreme response patterns. For this reason, item selection within each lexical category deliberately combined *high-frequency*, *mid-frequency*, and *low-frequency items* drawn from the coursebook corpus. This stratified approach was intended to reduce the likelihood of uniformly correct or uniformly incorrect responses and instead to elicit performance distributions that fall between ceiling and floor effects, allowing meaningful category-level and item-level patterns to emerge. The pilot results confirmed this expectation, yielding non-polarised response profiles that are suitable for exploratory analysis and further instrument development.

Two additional decisions were adopted deliberately for this pilot stage. Production responses were scored *binary* (correct/incorrect), with tolerance for surface errors that do not impede meaning, in order to prioritise clear and replicable coding and to indicate whether the instrument could reliably capture recognition–production asymmetries before introducing finer-grained scoring schemes. In addition, the pilot focused on written sentence production, which is straightforward to collect at scale, reveals collocational and argument-structure difficulties, and minimises logistical complexity in classroom settings. These choices reflect pragmatic pilot-stage trade-offs rather than claims of comprehensive measurement and are intended to inform subsequent refinement, including graded scoring, reliability estimation, and the incorporation of oral production tasks.

Test format and administration

Each test form comprised two sections that used the same 25 lexical targets in the same order. The first section assessed receptive knowledge: each target item was presented with a four-option multiple-choice question (a–d) asking for the meaning of the item. The second section assessed productive knowledge: participants were required to use each target item in a self-constructed sentence to demonstrate correct usage in context. Tests were implemented in Google Forms. During class, each group received access to its assigned test via a unique QR code projected on the interactive whiteboard; students scanned the code and completed the test on their devices.

Scoring and coding protocol

All gathered data were exported into a Microsoft Excel table and coded systemically. Since the focus is on receptive and productive skills, being able to correctly use the words in proper context, all errors, which did not impede the understanding of the sentence, as a whole, were ignored. Thus, all correct uses were marked as 1 and the incorrect uses marked as 0, for both open and closed questions.

Data analysis

The data analysis showed the results from the written tests of 186 students. Each student completed a test, with 25 closed questions and 25 open questions. The students completed a total of 9300 questions, which were then analyzed. The information was exported into Microsoft Excel and coded by hand individually.

In addition to descriptive accuracy comparisons, category-level standardised mean differences (Cohen's d) were calculated as an exploratory indicator of the magnitude of receptive–productive differences. For each lexical category, mean accuracy and standard deviation were computed by pooling all items belonging to that category. Effect sizes were calculated using a pooled estimate of variability across receptive and productive scores. These estimates are reported for descriptive and diagnostic purposes only and are not intended as confirmatory hypothesis tests or as evidence of full psychometric validation.

Rationale for design choices

Using nine parallel forms with non-overlapping target sets increased the lexical sampling from a single 25-item instrument to 225 unique lexical targets, improving the generalizability of findings across medical lexis. Selecting items stratified by corpus frequency (high, mid, low) made it possible to examine frequency effects on recognition and production as well as category effects.

Ethical considerations

Participation was voluntary and took place during regular instructional time. Students were informed of the study purpose and procedures and provided consent to use anonymized responses for the purposes of research.

Results

Overview of dataset and analytic focus

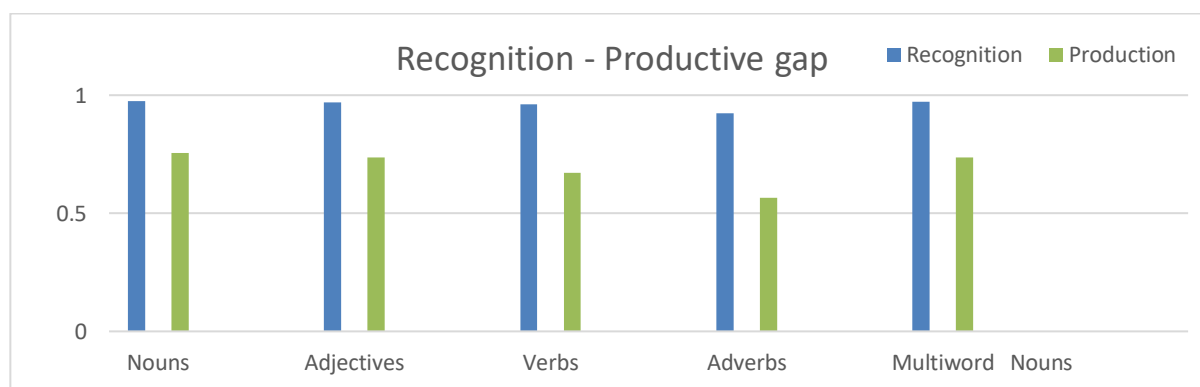
A total of 186 second-year medical students completed one of nine parallel MedLex-25 forms, yielding 9,300 item-level responses distributed across 225 unique lexical targets. The analysis below focuses on instrument performance in a pilot context, examining whether MedLex-25 can (a) elicit recognition–production asymmetries under classroom conditions, (b) differentiate performance across lexical categories, and (c) avoid ceiling or floor effects that would limit its usefulness as a formative diagnostic. Detailed item-level lists and full receptive and productive response distributions are reported in the Appendix; the Results section summarises the evidence these data provide about the behaviour and potential diagnostic utility of the instrument.

Evidence of receptive–productive asymmetry (stability of the core pattern)

Across the dataset, receptive accuracy exceeded productive accuracy for all lexical categories tested. Counts of perfectly recognized items (i.e., items with 100% correct receptive responses across groups) were: Nouns 33/45, Adjectives 27/45, Verbs 23/45, Adverbs 25/45, and Multiword items 28/45. In contrast, no lexical item achieved perfect production accuracy. The recurrence of this receptive–productive gap across five lexical categories and nine non-overlapping test forms suggests that MedLex-25 is capable of eliciting the intended asymmetric pattern in classroom conditions, rather than producing form-specific or random variation.

Figure 1

The Recognition – Production gap



In this pilot sample, MedLex-25 appears able to capture the recognition–production contrast it was designed to probe, providing preliminary evidence that the paired-task format can reveal productive weaknesses even when receptive familiarity is high.

Evidence of category-level differentiation

Performance patterns differed across lexical categories in ways broadly consistent with expectations from prior research. Nouns and multiword items tended to suggest smaller receptive–productive gaps, whereas verbs, adjectives, and adverbs exhibited larger declines from recognition to production. Adverbs, in particular, included items with both lower receptive values and very low production accuracy (e.g., *erroneously*, production = 0.071). While these results do not constitute definitive category ranking, they indicate that the instrument is sensitive enough to reflect category-related variation rather than yielding uniform performance across lexical types.

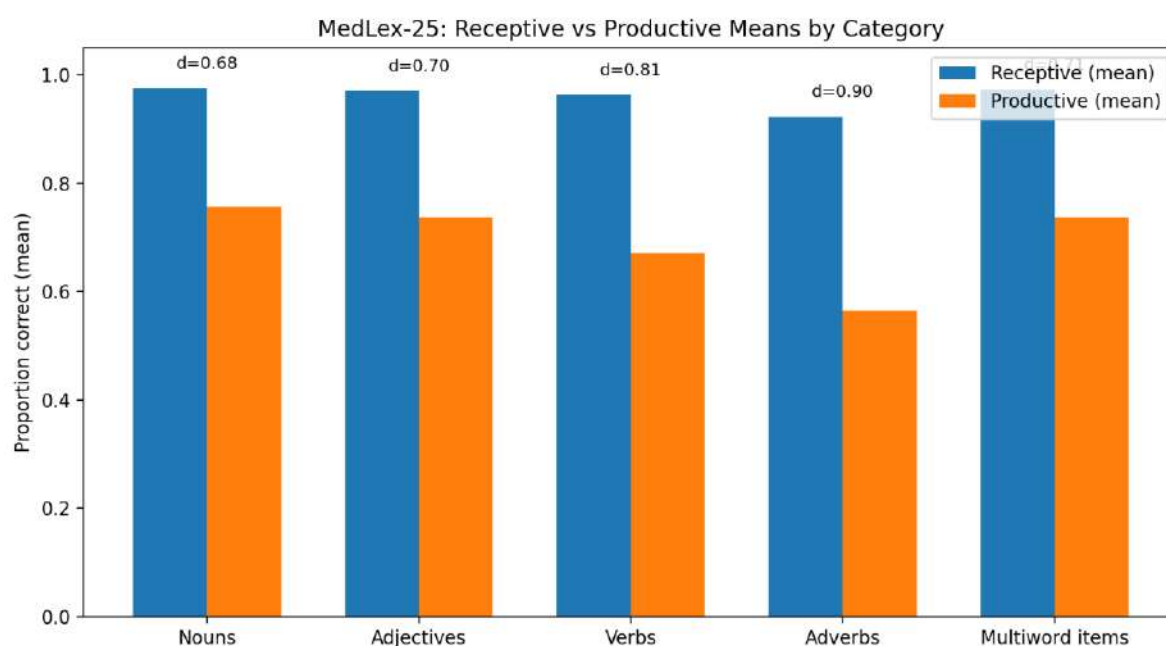
The category-stratified design of MedLex-25 appears to allow tentative differentiation among lexical categories, supporting its potential use for identifying broad areas of productive vulnerability in a formative, exploratory manner.

Magnitude of receptive–productive differences

To further characterise the magnitude of category-level differences, standardised mean differences (Cohen’s *d*) were calculated between receptive and productive scores. Effect sizes ranged from moderate to large across categories, with smaller gaps observed for nouns and multiword items and larger gaps for verbs and adverbs. These estimates quantify the descriptive patterns reported above and indicate that receptive–productive contrasts vary meaningfully across lexical categories in this pilot sample. Figure 2 illustrates mean receptive and productive accuracy by category alongside the corresponding effect-size estimates.

Figure 2

Mean receptive and productive accuracy by lexical category with corresponding category-level effect sizes (Cohen's d). Values are shown for descriptive purposes in a pilot context.



Absence of ceiling effects in production

Short classroom diagnostics risk becoming uninformative if performance clusters at ceiling or floor levels. In the present pilot, production accuracy did not reach ceiling for any item, and productive scores were distributed across a broad range within each lexical category. This pattern reflects the deliberate inclusion of high-, mid-, and low-frequency items and resulted in non-polarised response distributions that permit exploratory analysis.

The absence of production ceiling effects in this pilot indicates that MedLex-25 may offer sufficient dynamic range to identify relative productive difficulty among items, a prerequisite for meaningful classroom diagnosis, though this should be re-examined in larger and more diverse samples.

Item-level diagnostic interpretability

At the item level, MedLex-25 yielded response profiles that appear pedagogically interpretable. Some items showed low receptive and low productive accuracy, whereas others were recognized reliably but produced inconsistently. These contrasting patterns are visible across categories and point to different potential sources of difficulty (e.g.,

limited exposure versus difficulties with proceduralised use). The instrument does not explain these difficulties, but it makes them visible in a systematic way. Full item-level results are provided in the Appendix.

In this pilot application, MedLex-25 generates item-level information that could support instructional decision-making by helping instructors distinguish between items requiring additional exposure and those requiring focused production practice.

Feasibility of classroom administration

MedLex-25 was administered using Google Forms during regular class sessions across nine parallel versions. Completion rates were high, and binary coding of production responses seemed manageable at classroom scale. The resulting dataset provides preliminary evidence that the instrument can be implemented without specialised testing infrastructure or disruption to normal instruction.

While feasibility does not constitute validation, the successful administration and data yield in this pilot context indicate that MedLex-25 can be used in everyday classroom workflows, supporting its intended role as a formative diagnostic tool rather than a high-stakes assessment.

Discussion

Limits of pilot validation

The present study reports findings from an early-stage pilot application of MedLex-25, and its conclusions must be interpreted within clear methodological limits. First, no reliability coefficients (e.g., internal consistency or inter-rater reliability) were calculated at this stage, as the primary aim was to examine feasibility and diagnostic behaviour rather than measurement precision. Second, the analysis is descriptive and exploratory; no inferential statistical testing was conducted, and observed category-level patterns should therefore be treated as indicative rather than confirmatory. Third, the use of binary scoring for production responses, while appropriate for a pilot classroom diagnostic, necessarily reduces variance and obscures gradations of partial knowledge. Finally, the sample is context-specific, consisting of Bulgarian L1 medical students with prior exposure to Latin-based terminology, which limits the generalisability of the results to other learner populations. These constraints are acknowledged as inherent to the pilot phase and define the scope of the claims made in this study.

Diagnostic instrument

Within these limits, the pilot results allow a preliminary evaluation of MedLex-25 as a diagnostic instrument rather than as a lexical test per se. The paired receptive-productive format functioned as intended, consistently producing non-polarised response distributions and revealing recognition-production gaps across lexical categories. The instrument also appeared feasible to administer and score within routine classroom settings, supporting its intended formative use. At the same time, the pilot highlighted several aspects requiring further development: the absence of reliability estimates, the reliance on binary production scoring, and the restriction to written production constrain interpretability and measurement precision. These outcomes suggest that, while the current version of MedLex-25 is suitable for exploratory classroom diagnosis, it remains a work in progress that would benefit from expanded scoring schemes, reliability testing, and validation across learner populations and modalities.

The findings of this study provide preliminary evidence of a recurrent recognition-production gap in medical ESP vocabulary, with receptive knowledge consistently outperforming productive ability across all lexical categories. Although students demonstrated strong recognition of many items—33 out of 45 nouns, 27 out of 45 adjectives, 23 out of 45 verbs, 25 out of 45 adverbs, and 28 out of 45 multiword items achieved perfect receptive scores—no lexical item in the entire instrument was produced perfectly by all groups. This divergence points to a distinct bottleneck in learners' ability to retrieve and use known lexical items accurately in context, rather than a global lack of vocabulary knowledge.

The category-specific patterns reveal meaningful differences in how medical vocabulary is acquired and activated. Nouns and multiword items emerged as the most stable categories, with high recognition and relatively strong production accuracy. Their comparatively small recognition-production gap can be partly attributed to students' prior exposure to Latin and anatomy courses during their first year of medical studies, which likely strengthens their familiarity with technical terms of Latin or Greek origin. Many medical nouns overlap cross-linguistically between English, Bulgarian, and Latin, making them more transparent to learners. This helps explain why nouns such as

neutrophil or *emphysema* were recognized easily, whereas non-classical items such as *stool* (0.714 receptive accuracy) posed far greater difficulty. Even so, several noun items revealed notable weaknesses—*enanthem* (0.333 production) and *noncompliance* (0.500)—suggesting that transparent etymology alone does not secure productive mastery.

Verbs and adjectives displayed substantially larger disparities. Verb recognition was weaker overall, and verb production showed the steepest decline, with extremely low accuracy for items such as *underline* (0.111), *exhibit* (0.292), and *administer* (0.400). These patterns parallel long-standing observations in applied linguistics regarding the complexity of verb argument structure, voice alternations, and prepositional frames. Adjectives exhibited a similar, though slightly less severe, pattern: several were reasonably well recognized but remained difficult to produce accurately, including *concise* (0.350), *communicable* (0.435), and *innate* (0.529). The challenge here is primarily collocational and stylistic—learners may understand the meaning of an adjective but lack the intuitive sense of its typical noun pairings within medical discourse.

Adverbs proved to be the least stable category of all. Despite a relatively good number of perfectly recognized adverbs (25/45), several items showed weak receptive performance, including *erroneously* (0.286) and *considerably* (0.478). Production accuracy was markedly low, with *erroneously* (0.071) representing the weakest single score in the dataset. Unlike nouns, adverbs receive little to no reinforcement outside English classes, as they are absent from parallel instruction in Latin, anatomy, or other domain subjects. Their syntactic mobility and subtle semantic nuance further complicate real-time retrieval. These factors collectively help explain why adverb production—not noun production—emerges as the greatest vulnerability in students' lexical repertoires.

The item-level error patterns generated by the MedLex instrument offer practical diagnostic insight. Items with low receptive scores, such as *stool*, *erroneously*, or *anticipate*, likely require greater exposure, clearer definitional support, and repeated contextual encounters. By contrast, items with strong receptive scores but low production accuracy—*irrigate*, *concise*, *triggered by*, or *susceptible to*—represent cases where learners possess form–meaning knowledge but lack the procedural fluency to use these items in grammatically and pragmatically acceptable ways. These two error profiles

call for different pedagogical responses: the former demands enriched input, while the latter necessitates controlled output and scaffolded opportunities to practise syntactic integration.

Pedagogically, the results suggest that curricula should prioritise structured, production-oriented practice for verbs, adjectives, and adverbs, with special attention to argument structure, collocations, and canonical adverbial placement. Multiword items—though relatively strong—still benefit from reinforcement that consolidates their status as formulaic units while encouraging flexible use across varying sentence patterns. Retrieval-based tasks, spaced repetition, and feedback that simultaneously addresses meaning and form appear particularly promising for narrowing the recognition–production divide.

Several methodological considerations should temper generalisation. The binary scoring scheme for production simplifies gradations of partial knowledge and may obscure intermediate competence levels. Although nine non-overlapping test forms expanded lexical coverage, this design prevented any single learner from encountering all items, limiting within-participant depth. The sample was restricted to Bulgarian L1 medical students at a single institution with prior Latin exposure, which strongly shapes the category-specific results. Finally, the study assessed written production only; future research should include oral production and spontaneous interaction to capture real-time lexical retrieval demands.

Looking ahead, the findings point toward the need for a more granular, qualitative examination of why specific items fail at either the recognition or production stage. A future study should focus on a single lexical category—or even a single problematic subset within that category—and investigate the sources of difficulty through etymological analysis, cross-linguistic comparison, and interviews that probe students' prior exposure to terms in Latin, Bulgarian, and non-ESP coursework. Understanding the role of etymology, transparency, and disciplinary reinforcement may clarify why learners excel with classical-pattern nouns but struggle with everyday medical terms such as *stool*, or why adverbs remain fragile despite adequate receptive knowledge. Such work would deepen our understanding of the cognitive and pedagogical mechanisms driving lexical development in medical ESP and inform the refinement of diagnostic tools like MedLex.

Conclusion

The results of this pilot study provide initial evidence that MedLex-25 may elicit systematic recognition–production asymmetries and yield item-level diagnostics that could be pedagogically useful in a classroom context. Rather than presenting a completed, standardised instrument, the study indicates that a compact, corpus-sourced, paired receptive–productive design may end up being a promising approach for formative diagnosis of medical English vocabulary.

These findings should be read as exploratory and context-bound: they suggest where instructors might prioritise production practice (notably verbs, adjectives and adverbs in this sample) and they provide practical evidence about the instrument’s feasibility under supervised classroom conditions. At the same time, the results point to limitations inherent in this early stage of development — limited generalisability from a single L1 group, binary scoring choices that reduce variance, and the absence here of full reliability and inferential validation.

Accordingly, MedLex-25 is best understood as a working diagnostic tool that provides actionable, provisional information for teachers and that furnishes a basis for further psychometric refinement. Future work should expand sampling, introduce graded scoring and inter-rater reliability checks, report reliability coefficients and effect sizes, and attempt replication across different instructional contexts and L1 backgrounds to establish broader generalisability. These steps will help move the instrument from formative piloting toward more consistent validation while preserving its practical orientation to classroom diagnosis.

References

- Amanda Edmonds, Jon Clenton, & Hosam Elmetaher. (2022). *Exploring the construct validity of tests used to assess L2 productive vocabulary knowledge*. System, 108, Article 102855. <https://doi.org/10.1016/j.system.2022.102855>
- Begagić, M. (2025). English language students’ productive and receptive knowledge of collocations. *ExELL*, 2(1), 46-67. <https://exell.untz.ba/index.php/exell/article/view/37>
- Boone, G., De Wilde, V., & Eyckmans, J. (2023). A longitudinal study into learners’ productive collocation knowledge in L2 German and factors affecting the

- learning. *Studies in Second Language Acquisition*, 45(2), 503-525.
<https://doi.org/10.1017/S0272263122000377>
- Doykova, Ilina. (2017). Verb Choice in Medical Research Articles (ESP). *Studies in Linguistics, Culture, and FLT*, 2, 240-252.
<https://doi.org/10.46687/SILC.2017.v02.020>
- Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics*, 19(4), 443-477. <https://doi.org/10.1075/ijcl.19.4.01dur>
- Durović, Z., & Bauk, S. (2022). Testing corpus linguistics methods in ESP vocabulary teaching. *The Journal of Teaching English for Specific and Academic Purposes*, 10(2), 319-337. <https://doi.org/10.22190/JTESAP2202319D>
- Fang, L., Ma, Q., & Yan, J. (2021). The effectiveness of corpus-based training on collocation use in L2 writing for Chinese senior secondary school students. *Journal of China Computer-Assisted Language Learning*, 1(1), 80-109.
<https://doi.org/10.1515/jccall-2021-2004>
- Fraser, S., Higa, M. K., & Davies, W. (2025). Delivering an ESP pedagogic word list: Integrating corpus analysis, materials design, and software development. *Languages*, 10(3), 46. <https://doi.org/10.3390/languages10030046>
- Kang, Y., & Niu, M. (2024). Enhancing language proficiency in medical English education through multimodal corpus integration. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, 19(1), 1-1.
<https://doi.org/10.4018/IJHISI.356368>
- Kokkinakis, D. (2006). Collection, encoding and linguistic processing of a Swedish medical corpus – The MEDLEX experience. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA).
- Lee, S. (2025). The relationship between receptive and productive knowledge of L2 English collocations. *International Journal of Applied Linguistics*, 35, 109-133.
<https://doi.org/10.1111/ijal.12605>
- Malmström, H., Pecorari, D., & Warnby, M. (2025). Teachers' receptive and productive vocabulary sizes in English-medium instruction. *Journal of Multilingual and Multicultural Development*, 46(7), 1905-1923.
<https://doi.org/10.1080/01434632.2023.2260781>
- Miller, V., & Larsson, T. (2024). The effect of linguistic and extralinguistic features on EFL adverb placement: A partial replication study of Larsson et al. (2020). *International Journal of Learner Corpus Research*, 10(2), 338-364.
<https://doi.org/10.1075/ijlcr.00046.ozk>

- Mohammadi, M., Valizadeh, M., Zohdi Jalal, P., & Xodabande, I. (2024). University students' academic vocabulary development through mobile-assisted learning: Exploring the impacts on receptive and productive knowledge. *Heliyon*, 10(7), Article e28103. <https://doi.org/10.1016/j.heliyon.2024.e28103>
- Nakamura, C., Arai, M., Hirose, Y., & Flynn, S. (2022). L2 learners do not ignore verb's subcategorization information in real-time syntactic processing. *Frontiers in Psychology*, 12, Article 689137. <https://doi.org/10.3389/fpsyg.2021.689137>
- Nguyen, N. G. H. (2025). Challenges and strategies in teaching English for medical purposes: A literature review. *International Journal of Language Instruction*, 4(3), 57-73. <https://doi.org/10.54855/ijli.25434>
- Pu, P., Chang, D. Y.-S., & Wang, S. (2024). Incidental learning of collocations through different multimodal input: The role of learners' initial L2 proficiency. *System*, 125, Article 103416. <https://doi.org/10.1016/j.system.2024.103416>
- Rogers, J., (2025). Repetition, Retrieval and Spaced Practice In C.A.Chapelle, D.Trenkic & A.De Bruin (Eds.), *The Encyclopedia of Applied Linguistics (Second Language Acquisition and Bilingual Processing)* (pp.1-6). John Wiley & Sons <https://doi.org/10.1002/9781405198431.wbeal20349>
- Sonbul, S., El-Dakhs, D. A. S., & Masrai, A. (2023). Second language productive knowledge of collocations: Does knowledge of individual words matter? *Studies in Second Language Acquisition*, 45(2), 480-502. <https://doi.org/10.1017/S0272263122000341>
- Tran, H., & Waluyo, B. (2021). Receptive knowledge of nouns and collocations and the impact on English skill performances. *GEMA Online Journal of Language Studies*, 21, 124-147. <https://doi.org/10.17576/gema-2021-2101-08>
- Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, 30(1), 79-95. <https://doi.org/10.1017/S0272263108080042>
- Yamagata, S., Nakata, T., & Rogers, J. (2023). Effects of distributed practice on the acquisition of verb–noun collocations. *Studies in Second Language Acquisition*, 45(2), 291-317. <https://doi.org/10.1017/S0272263122000225>
- Zhang, X., & Sukying, A. (2021). Receptive and productive knowledge of lexical collocations in Thai university learners of English. *European Journal of English Language Teaching*, 6(6). 266-285. <https://doi.org/10.46827/ejel.v6i6.4067>

Reviewers:

1. Anonymous
2. Anonymous

Handling Editor:

Stan Bogdanov, PhD
New Bulgarian University