# AI-медиирана трансформация на входа в медицинския английски: разпознаване на реч и надеждност на транскрипцията

Евгени Станчев

# AI-Mediated Input Transformation in Medical English: Speech Recognition and Transcript Reliability

Evgeni Stanchev

**Abstract:**

This study evaluates automatic speech recognition (ASR) performance on a controlled text-to-speech (TTS) rendering of a 2,692-token biochemistry passage used in Medical English instruction. Synthetic audio was transcribed by four online ASR platforms: Turboscribe, Riverside, 1transcribe, and Smartnote AI. Two freely accessible services (Turboscribe and Riverside) produced near-complete transcripts with 88.0–88.2% coverage and moderate Word Error Rates (WER) of 15.23% and 15.02%, respectively. Two paid services returned partial outputs covering 30.1% and 54.6% of the reference text, yielding WER values of 72.14% and 47.77%, primarily driven by missing segments rather than dense substitution errors. Across systems, deletions were the dominant error type. In full transcripts, deletion rates were approximately 12.6% of reference tokens, with roughly 22% of deletions involving structural markers such as figure references and numeric labels. Meaning-altering semantic substitutions were rare (<1% of tokens) but pedagogically significant (e.g., *imino* → *amino*, *cystine* → *cysteine*, *pH* → *phase*). The findings demonstrate that coverage must be reported alongside WER in ESP contexts, that full transcript export is essential for pedagogical reliability, and that lightweight human verification is required where terminological precision is critical.

**Keywords:** Automatic Speech Recognition, Text-to-Speech, Medical English, English for Specific Purposes

**For contacts:** Evgeni Stanchev, Lecturer and PhD Candidate, Medical University – Varna (MU Varna), evgeni.stanchev@mu-varna.bg

## INTRODUCTION

Text-to-speech and automatic speech recognition technologies are increasingly integrated into language instruction, offering consistent auditory input and new avenues for automated feedback. In English for Specific Purposes (ESP) settings, where learners encounter specialized vocabulary and tightly specified propositional content, controlled auditory models and reliable transcripts are particularly valuable for lesson design, assessment, and autonomous study (1, Avrianti et al., 2025; 2, El Fakir, 2025). ASR systems are now used not only for generating transcripts but also for enabling automated dictation tasks, pronunciation feedback, and interaction logging, which can support formative learning cycles and facilitate large-scale materials preparation (3, Sun, 2023). However, a service's practical utility in pedagogy depends on the reliability of its transcriptions. Word Error Rate (WER) is the standard metric for evaluating ASR outputs, but WER alone is misleading when transcript coverage differs between services; insight into pedagogical impact requires both coverage and error composition (4, Park et al., 2024; 5, Wikipedia, 2024). This study therefore assesses ASR outputs from a controlled TTS source using token-level alignment and qualitative inspection to determine which kinds of errors are pedagogically tolerable and which require intervention.

ASR outputs are generated through probabilistic modeling, where lexical selection is based on acoustic likelihood and language model prediction rather than semantic

understanding (7, Jurafsky & Martin, 2023). As a result, transcription becomes a predictive reconstruction rather than a deterministic rendering of speech. This distinction is especially relevant in ESP environments, where lexical precision encodes conceptual specificity. Research on algorithmic mediation further suggests that automated systems restructure communication flows by embedding statistical assumptions into output representations (8, Gillespie, 2014).

This raises a critical pedagogical question: at what point does transcription error become conceptual distortion? In general language learning, minor deletions or function-word omissions may not significantly impede comprehension. In medical and biochemical discourse, however, minimal lexical variation—such as the distinction between cystine and cysteine—can represent distinct biochemical realities. Evaluating ASR reliability in ESP settings therefore requires attention not only to quantitative error rates but to epistemic integrity: whether the knowledge structure encoded in the transcript remains intact.

## METHODS

The reference text was a 2,692-token excerpt from a biochemistry coursebook chapter on proteins (Harvey & Ferrier). The passage contains domain terminology (amino acids, disulfide bonds), structural markers (figure references, page numbers), numerals, and chemical nomenclature. A single TTS rendering was produced using **NaturalReader** (female Canadian voice) and recorded in **Audacity** at 44.1 kHz, 16-bit PCM, mono, producing a 15:27 (minutes: seconds) audio file used as the shared input for all ASR services. Using TTS eliminated speaker variability so that observed differences reflect ASR behavior rather than human prosodic or disfluency factors.

Four online ASR platforms were evaluated: **Turboscribe**, **Riverside**, **1transcribe**, and **Smartnote AI**. Turboscribe and Riverside were used under freely accessible plans and returned near-complete transcripts. 1transcribe and Smartnote AI were evaluated through paid-access plans but provided only partial or preview outputs during testing. Each service received the identical audio file; where a platform returned only a preview, that preview was retained verbatim and coverage was computed to quantify how much of the reference each hypothesis represented.

Reference and hypothesis texts were normalized before alignment to focus matching on lexical content: all characters were lowercased, sentence punctuation removed, multiple whitespaces collapsed, and obvious symbols mapped to text. Orthographic variants were inspected post hoc and reported where relevant rather than normalized automatically. A Levenshtein dynamic program aligned token sequences to produce counts of deletions (D), substitutions (S), and insertions (I). For example, Turboscribe alignment yielded D=339, S=66, I=5 (N=2,692), while Riverside produced D=341, S=42, I=22 (N=2,692). Partial outputs from 1transcribe (D=1,925; S=31; I=15; N=2,692) and Smartnote AI (D=1,253; S=41; I=11; N=2,692) reflected truncated transcript returns.WER was computed as (D + S + I) / N, where N is the reference token count (2,692), following the standard formulation (4, Park et al., 2024; 5, Wikipedia, 2024). Coverage was defined as hypothesis_word_count / N. Coverage values ranged from 88.0–88.2% for full transcripts to 30.1% and 54.6% for partial outputs. Following automated alignment, hypotheses were manually inspected to classify pedagogically relevant error types: structural omissions, function-word deletions, orthographic variants, and semantic substitutions.

## RESULTS

The two freely accessible platforms, **Turboscribe** and **Riverside**, produced transcripts with approximately 88% coverage and WER values around 15%, driven primarily by deletions and a smaller number of substitutions and insertions. In contrast, **1transcribe** and **Smartnote AI** returned partial outputs, with coverage levels of approximately 30% and 55%, respectively. Their WER values were substantially higher (approximately 47–72%), reflecting the large number of missing tokens in the returned transcripts.

Table 1 summarizes reference length (N), hypothesis length, coverage, deletion (D), substitution (S), insertion (I) counts, and Word Error Rate (WER) for all four ASR systems evaluated.

| Service | Ref N | Hypothesis | Coverage | D | S | I | WER |
|---|---|---|---|---|---|---|---|
| Turboscribe | 2,692 | 2,368 | **87.96%** | 339 | 66 | 5 | **15.23%** |
| Riverside | 2,692 | 2,378 | **88.34%** | 341 | 42 | 22 | **15.00%** |
| 1transcribe | 2,692 | 822 | **32.76%** | 1,925 | 31 | 15 | **73.21%** |
| Smartnote AI | 2,692 | 1,490 | **55.35%** | 1,253 | 41 | 11 | **48.77%** |

*Table 1. Coverage and Word Error Rate (WER[9]) for Four ASR Systems on a 2,692-Token Biochemistry Passage*

Across all four platforms, deletions constituted the most frequent error category. In the higher-coverage outputs from Turboscribe and Riverside, deletion counts exceeded substitution counts, indicating that uncertain tokens were more often omitted than replaced. In the partial outputs from 1transcribe and Smartnote AI, deletions were amplified by truncated segments, which significantly increased overall error totals.

Deletion errors clustered around structural markers and small function words. Tokens such as figure references, page numbers, articles, and prepositions were frequently omitted. While these omissions reduce navigational usability and may affect alignment with visual materials, they generally preserve the core propositional content of the passage.

By contrast, semantic substitutions—although less frequent—posed greater pedagogical risk. Examples identified during manual inspection included imino → amino, cystine → cysteine, protonated → protonatid, and pH → phase. These substitutions alter domain meaning and could mislead learners if left uncorrected, particularly in contexts where terminological precision is essential.

The elevated WER values observed for 1transcribe and Smartnote AI were primarily attributable to reduced transcript coverage rather than dense substitution within returned segments. When only preview portions are available, computed WER conflates incomplete transcript availability with transcription accuracy. For this reason, reporting coverage alongside WER is essential for transparent and meaningful comparison across ASR platforms.

## DISCUSSION

[9] Note. WER = (D + S + I) / N. Coverage = hypothesis word count / reference word count. Two services (1transcribe and Smartnote AI) returned partial transcripts; WER for these outputs primarily reflects reduced coverage.

The controlled TTS→ASR pipeline demonstrates three practical implications for ESP practitioners. First, coverage must be reported together with WER, as recommended in recent work on WER estimation (4, Park et al., 2024). Without coverage, comparisons may misrepresent system reliability. Second, deletion errors should be interpreted in relation to pedagogical impact. Structural token loss may inconvenience navigation, whereas semantic substitutions directly affect conceptual integrity and require prioritization. Third, full transcript export is necessary for classroom materials and corpus building; preview-only services should be treated as diagnostic tools rather than as production-level transcription systems.

From a methodological perspective, the study confirms that WER remains a useful baseline metric when interpreted with contextual awareness (5, Wikipedia, 2024). For ESP contexts in medical education, domain adaptation and targeted post-editing are advisable to reduce the likelihood of meaning-altering substitutions, consistent with broader findings on ASR use in language learning (3, Sun, 2023).

## PEDAGOGICAL IMPLICATIONS FOR MEDICAL ENGLISH INSTRUCTION

In medical English classrooms, transcripts are frequently used to support note-taking, reinforce listening comprehension, and scaffold domain terminology acquisition. A practical implementation model would involve a three-stage workflow: (1) generate TTS or recorded lecture input; (2) produce a full-export ASR transcript; (3) conduct targeted verification of domain-specific terminology before classroom use. This approach preserves efficiency while maintaining terminological integrity. In advanced ESP contexts, instructors may even incorporate ASR error analysis as a metalinguistic activity, encouraging learners to identify discrepancies between transcript and audio as a way of reinforcing precise lexical distinctions.

By repositioning ASR output as a pedagogical artifact rather than a definitive textual authority, educators can both harness technological efficiency and cultivate critical digital literacy among students.

## CONCLUSION

In this comparison, two accessible free ASR services produced near-complete transcripts with moderate WER values, while two paid services supplied partial outputs that inflated WER through low coverage. Deletions were the most frequent error type and often affected structural tokens, whereas infrequent semantic substitutions posed the highest pedagogical risk. For ESP applications in medical and healthcare education, practitioners should require full transcript access, report coverage alongside WER, and implement targeted human verification or domain adaptation to protect terminological accuracy.

## REFERENCES

1. Avrianti, N., Yuliana, Y. G. S., & Susilawati, E. (2025). Assessing the Effectiveness of Text-to-Speech Applications on Enhancing Listening Comprehension and Student Engagement. *International Journal of Research in Social Sciences*.

2. El Fakir, Z. (2025). TTS and STT in Service of Education. In *Educational Technologies and Learning Ecosystems*. MDPI.

3. Sun, W. (2023). The Impact of Automatic Speech Recognition Technology on Second Language Pronunciation and Speaking Skills of EFL Learners: A Mixed Methods Investigation. *Frontiers in Psychology*, 14, 1210187.

4. Park, C., Chen, M., & Hain, T. (2024). Automatic Speech Recognition System-Independent Word Error Rate Estimation. *LREC/COLING 2024*.

5. Wikipedia Contributors. (2024, *Word error rate*). Wikipedia. *WER metric derived from Levenshtein distance; used widely in ASR evaluation.* Retrieved from https://en.wikipedia.org/wiki/Word_error_rate

6. Verbeek, P. P. (2005). What Things Do: Philosophical Reflections on Technology, Agency, and Design. Pennsylvania State University Press.

7. Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed. draft). Stanford University.

8. Gillespie, T. (2014). The Relevance of Algorithms. In T. Gillespie, P. Boczkowski, & K. Foot (Eds.), Media Technologies: Essays on Communication, Materiality, and Society. MIT Press.