Applications of Language Studies
jals.hmu.ac.ir

Identifying Discourse Features for Medical English Instruction: Evidence from YouTube-Based Corpus Analysis

Ebrahim Samani[1], Abdolhossein Heydari[2], Hasan Tarlanialiabadi[3]

1. Assistant Professor, Department of English Language Teaching, Farhangian University, Tehran, Iran
e.samani@cfu.ac.ir (*corresponding author*)

2. Assistant Professor, Department of English Language Teaching, Farhangian University, Tehran, Iran
a.heidari@cfu.ac.ir

3. Assistant Professor, Department of English Language Teaching, Farhangian University, Tehran, Iran
htarlani1983@cfu.ac.ir

Abstract

Voice recognition and text-mining tools have democratized corpus construction, allowing educators to create responsive materials based on authentic discourse trends. In global healthcare contexts, where English functions as a lingua franca, cultivating discourse competence is essential for safe and effective communication. This study employed corpus linguistics methods to analyze authentic medical English sourced from YouTube videos. A specialized corpus was compiled and analyzed using *Otter.ai* and *Sketch Engine* to examine linguistic patterns, collocations, and discourse structures in real-world medical communication. Findings revealed that medical professionals frequently employ modal verbs (e.g., *can, should, must*), passive constructions, and high-frequency collocations (e.g., *blood pressure, heart rate* and *patient care*) that serve distinct pragmatic functions. Moreover, interactional phrases such as *let me explain* and *can you see* and discourse markers such as *so* and *now* are pervasive, reflecting the interpersonal and procedural nature of clinical discourse. Cross-cultural comparison further indicated notable variation in communication style, with Western contexts showing greater empathy and shared decision-making, and South Asian contexts favoring directive, authoritative tones. These patterns provide insight into the pragmatic and intercultural dimensions of medical English and highlight the potential of YouTube-based corpora as authentic resources for English for Medical Purposes (EMP) instruction.

Keywords: YouTube, discourse analysis, corpus linguistics, ESP, Sketch Engine, Otter.ai

Citation: Samani, E., Heydari, A., & Tarlanialiabadi, H. (2026). Identifying discourse features for medical English instruction: Evidence from YouTube-based corpus analysis. *Applications of Language Studies*, *4*(1), 89–113. https://doi.org/10.22034/jals.2026.2068356.1101

Received: 2025-08-08
Revised: 2025-11-30
Accepted: 2026-01-27
Published online: 2026-02-15

P-ISSN: 2588-6975
E-ISSN: 2588-6983
Article type: Research Paper
DOI: 10.22034/jals.2026.2068356.1101

1. Introduction

In the context of English for Academic Purposes (EAP) and English for Occupational Purposes (EOP), medical English sits at a critical juncture where communication effectiveness can influence patient safety and institutional efficiency. It is not enough to teach isolated vocabulary; learners must internalize the discourse practices that are contextually appropriate for medical interviews, diagnoses, and reporting. The rise of evidence-based language teaching underscores the importance of using real data to inform what is taught in classrooms and how it is delivered. Thus, YouTube as a medium provides not only linguistic input but also audiovisual cues that mirror real-life exchanges.

The integration of technology in language teaching also aligns with the broader shift toward blended and autonomous learning. Medical professionals often engage in continuing education amid demanding schedules, and corpus-based learning via YouTube enables flexible, learner-centered approaches (Godwin-Jones, 2018; Jiang & Peng, 2025). Students can independently explore interactional norms and pragmatic features, thereby accelerating their acquisition of communicative competence beyond the boundaries of the classroom.

Recent work in English for Specific Purposes (ESP) pedagogy emphasizes the importance of multimodal input (Dooly & O'Dowd, 2018; Flowerdew & Wang, 2015) and learner engagement (Boulton, 2017; Hyland, 2019). Corpus consultation, when combined with collaborative tasks, helps foster discovery learning. These approaches have proven particularly effective in teaching medical interactional routines.

Effective communication is paramount in medical settings, influencing patient outcomes, clinical decision-making, and inter-professional collaboration. Medical professionals must develop high proficiency not only in technical terminology but also in the pragmatic and discourse-specific aspects of medical English. Traditional educational approaches, often reliant on textbooks and simulated scenarios, may inadequately reflect authentic communication encountered in real medical environments (Hafner & Miller, 2011; Tribble, 2015).

Recent technological advancements and the proliferation of digital resources offer innovative opportunities for language learning. Platforms like YouTube, which provide abundant, authentic, and varied medical-related content, have emerged as valuable resources for language acquisition. These real-world interactions captured in video form present students with genuine examples of how medical professionals communicate effectively within clinical, academic, and interpersonal contexts.

Corpus linguistics, an analytical approach focusing on studying language usage in authentic contexts, can effectively capture these realistic linguistic dynamics. Building and analyzing specialized corpora allows educators to identify and teach language patterns directly applicable to students' practical needs. By examining frequency, collocation, and discourse patterns, corpus-based analyses offer significant insights into linguistic competence development (Boulton, 2017; Hyland, 2008).

This research addresses the need for a deeper understanding of medical discourse competence by creating and analyzing a corpus sourced from medical-focused YouTube content. The analysis specifically targets discourse structures, common vocabulary, and pragmatic expressions essential for medical students and professionals. Through this study, the applicability of corpus-based insights for curriculum enhancement and language teaching strategies is explored, contributing to improved discourse competence among medical practitioners.

2. Literature Review

A particularly influential framework in discourse analysis has been the work of Swales (1990) on genre analysis, which introduced the CARS (Create a Research Space) model for academic writing. In the context of medical English, genre analysis can be applied to various communicative events, such as case presentations, patient handovers, and diagnostic reports. These genre structures have predictable moves and linguistic markers that can be taught effectively using corpus-informed methods.

Other researchers have expanded on the pragmatic aspect of discourse competence in medical contexts. For example, Bardovi-Harlig and Hartford (1993) explored the pragmatic

failures that occur when learners do not understand implicature or indirect speech acts — phenomena common in doctor-patient interactions. Pragmatic competence, therefore, becomes a critical component of discourse competence in high-stakes fields such as healthcare.

The role of lexical bundles — frequent recurring sequences of words — has also received increasing attention in medical ESP. Chen and Baker (2010) analyzed lexical bundles in professional academic writing and found discipline-specific patterns that are often underrepresented in instructional materials. In the medical domain, bundles such as "the patient was admitted," "in the case of," or "the results suggest that" signal expertise and fluency and can be highlighted using corpus tools like Sketch Engine.

In addition to lexical and pragmatic aspects, researchers such as Hyland (2004) have drawn attention to stance and engagement features — the ways in which speakers and writers express certainty, caution, or seek rapport with their audience. These features are particularly important in medical settings where conveying authority while maintaining empathy is a delicate balance. Identifying these markers through corpus tools enables targeted teaching of interpersonal communication skills.

Furthermore, the advent of learner corpora — databases built from language produced by learners — has created new opportunities to contrast native-like use with non-native patterns. Studies comparing expert and novice discourse (e.g., Granger et al., 2015) provide insights into typical learner errors, genre missteps, and areas where instruction can be most impactful. Incorporating learner corpora alongside expert models in medical English teaching helps bridge the gap between theory and practice.

Advancements in multimodal corpus linguistics also allow researchers to include visual, auditory, and paralinguistic features in their analyses. As noted by Baldry and Thibault (2006), meaning is constructed not just through words but through tone, gesture, and layout—an insight critical to medical communication where non-verbal cues often guide interpretation. YouTube videos, rich in such multimodal input, serve as ideal sources for this type of corpus work, making them highly relevant for modern ESP instruction.

Recent developments have also emphasized the socio-cultural dimensions of discourse competence, especially in cross-cultural healthcare environments where misunderstandings can have critical consequences. Studies by Roberts and Sarangi (2005) and Sarangi and Roberts (1999) highlighted how power dynamics, politeness strategies, and cultural expectations shape medical discourse. Incorporating such findings into ESP syllabi ensures that learners not only master terminology but also develop intercultural sensitivity.

Moreover, the use of spoken corpora in medical education has been advanced by works like Adolphs and Carter (2013), who demonstrated that spoken corpus analysis provides insights into real-life interactional patterns that are often invisible in written texts. Spoken corpora, particularly when multimodal, allow for the teaching of paralinguistic features such as intonation, pauses, and gestures, further bridging the gap between linguistic theory and communicative competence.

Discourse competence, a component of communicative competence proposed by Canale and Swain (1980), refers to the ability to produce and interpret connected and coherent speech or written texts. In medical contexts, this competence becomes vital, influencing diagnostic reasoning, patient interaction, and the conveyance of complex information (Chan, 2018). Scholars such as Candlin and Candlin (2002) have emphasized the high-stakes nature of medical communication, which demands not only mastery of terminology but also nuanced pragmatic skills.

Another vital contribution to the field is the concept of discourse communities introduced by John Swales (1990). In medical contexts, each sub-specialty (e.g., cardiology, pediatrics, radiology) may represent a distinct discourse community with its own norms, vocabulary, and genres. Corpus analysis allows educators to compare and contrast these communities, thereby designing syllabi that are both generalizable and specialized.

ESP, especially Medical English, has gained attraction as a distinct field within applied linguistics. Hutchinson and Waters (1987) identify the importance of needs analysis in ESP course design. In medical English, this involves tailoring content to reflect authentic tasks and scenarios encountered in healthcare settings. Basturkmen (2010) further highlights that effective ESP instruction must incorporate realistic discourse samples, something often

missing in traditional textbooks. From a pedagogical standpoint, Tribble (2015) emphasizes the need for task authenticity in ESP instruction, encouraging instructors to move beyond decontextualized grammar drills toward data-driven tasks. Such tasks include analyzing concordance lines, identifying collocations in authentic texts, and rewriting medical instructions to meet patient comprehension levels. These types of activities foster learner autonomy and deepen understanding of how language functions in real clinical settings.

With the advancement of digital media, platforms like YouTube have emerged as abundant sources of authentic discourse. Nguyen and Boers (2021) demonstrated that learners exposed to YouTube videos improved their pragmatic awareness and contextual language use. Such media offer access to genuine speech patterns, interactional features, and prosodic cues that are otherwise absent from scripted materials.

Corpus linguistics provides robust methodologies for analyzing language in context. According to Biber et al. (1998), corpora offer empirical bases for identifying recurrent patterns, collocations, and syntactic structures. Flowerdew (2004) advocates for corpus-informed ESP teaching, showing how frequency data and concordance lines could help learners internalize target language forms. The use of Sketch Engine, as explained by Kilgarriff et al. (2014), facilitates comprehensive corpus analysis by offering tools to examine frequency, collocation strength, and keyword patterns.

The proliferation of open-access corpora and tools like AntConc and LancsBox democratizes access to corpus linguistics, allowing even novice teachers or resource-constrained institutions to incorporate corpus-based learning. When combined with YouTube's vast, multimodal content, the potential for contextualized, low-cost, and learner-centered ESP instruction becomes limitless.

Previous corpus-based studies in medical English, such as those by Wang (2010) and Tseng (2013), have analyzed medical journals, textbooks, and classroom discourse. However, few have utilized dynamic and multimedia sources like YouTube, which can provide context-rich and dialogic language samples. These interactions are especially valuable for discourse-level learning because they capture spontaneous, real-time use of language across diverse medical scenarios.

The current study builds on these foundations, integrating transcription software (Otter.AI) and corpus analysis tools (Sketch Engine) to extract pedagogically useful patterns from YouTube-based medical content. This research aligns with the work of Csomay and Petrovic (2012), who advocated for multimedia corpora in educational contexts to bridge the gap between theoretical knowledge and applied communication.

Overall, the literature confirms the necessity of authentic input, digital integration, and corpus-driven methodology for effective medical English instruction. However, despite increasing interest in corpus-informed ESP pedagogy, few studies have systematically explored authentic, spoken medical discourse drawn from publicly available online sources. Existing research has primarily analyzed written texts, textbooks, or classroom simulations, which do not fully capture the spontaneous, interactional, and culturally diverse nature of real clinical communication. While the present study focuses on verbal transcriptions rather than multimodal features such as gesture, prosody, or gaze, YouTube-based data nonetheless provide access to naturally occurring, contextually grounded communication between medical professionals and lay audiences. Consequently, there remains a clear gap in understanding how naturally occurring medical discourse can inform the design of data-driven, learner-centered materials for English for Medical Purposes (EMP). Addressing this gap, the present study constructs and analyzes a YouTube-based medical English corpus to identify discourse patterns that can guide more authentic and contextually grounded language instruction.

Building on the reviewed literature and methodological framework, the current study addresses the following research questions:

1. How can a YouTube-based medical corpus be used to identify core discourse patterns that are essential for clinical communication and medical English instruction?

2. What are the dominant collocational, pragmatic, and genre-based features in the medical corpus?

3. Methodology

This study employed a corpus-based methodology to analyze authentic medical discourse extracted from YouTube videos. The research design included corpus construction,

transcription, cleaning, and analysis using digital tools tailored for linguistic research. The corpus analysis was grounded in a corpus-driven discourse analytic framework (Baker, 2006; Tognini-Bonelli, 2001), which emphasizes identifying recurring linguistic patterns emerging from authentic data rather than imposing pre-established categories. This approach was complemented by Hyland's (2005, 2019) model of stance and engagement, allowing the study to interpret recurrent lexical bundles, discourse markers, and pragmatic phrases in relation to their communicative functions within medical contexts. This dual framework ensured that both quantitative frequency patterns and qualitative discourse functions were systematically examined.

A total of 300-hour-long YouTube videos were selected based on the following criteria: (1) relevance to medical education or professional practice, (2) availability of spoken interaction (e.g., doctor-patient dialogues, lectures, interviews), and (3) diverse representation of English accents and medical specializations. Topics were related to medical fields including but not limited to medical school, microbiology, pharmacology, cardiology, and general healthcare. The video data were collected from YouTube through keyword-based searches using terms such as "medical communication," "doctor–patient interaction," and "health education." The search results were manually reviewed, and additional similar or active channels were explored through YouTube's suggested and related links. Channels consistently publishing medical communication content were selected, and relevant videos were downloaded regardless of their length or upload date, provided that they matched the study's focus. All videos were transcribed verbatim using Otter.ai, and the transcripts were reviewed and cleaned to remove noise or irrelevant segments. Table 1 indicates several examples of these channels:

As it is clear from Table 1, the videos that were downloaded from these websites were uploaded to the Otter.ai website. As Figure 1 shows, the selected videos were transcribed using Otter.AI, a high-accuracy speech recognition tool.

According to Figure 1 and the following transcription step, manual review was conducted to correct transcription errors and to ensure consistency in formatting, punctuation,

and speaker tagging. Non-verbal cues such as laughter, hesitation, or intonation were noted when relevant to discourse structure.

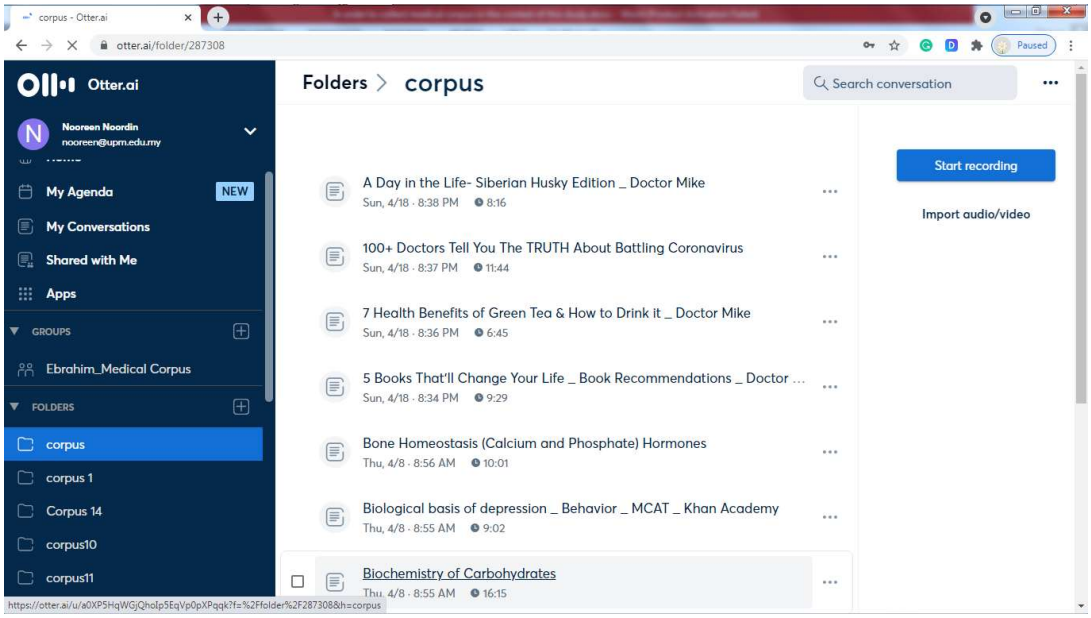
Table 1

Some of the Most Visited YouTube Channels Used for Compiling Medical YouTube Corpus in This Study

#	channels				
1	Med school insider	7	Dr Najeeb lectures	13	Doctor Mike
2	Registered nurse rn	8	Chubbyemu	14	Khan Academy health
3	Kati Morton	9	Armando Hasudungan	15	Osmosis
4	Nucleus medical media	10	FreeMedEd	16	One-minute medical school
5	Harvard medical school	11	Interactive biology		
6	Medcram videos	12	TED Talks		

Figure 1

A Snapshot of the Dashboard Page of the Otter.Ai Website



The transcripts were formatted into plain text files and annotated with basic metadata, including speaker roles (e.g., doctor, patient, educator), video title, and medical domain. Unrelated fillers and background noise were removed to maintain linguistic focus. The final corpus comprised approximately 3 million words, representing about 18,000 minutes (≈ 300 hours) of transcribed video content collected and processed over a three-month period using Otter.ai. The corpus annotation process was further refined by applying speech act categorization. Utterances were labeled based on whether they served as requests, confirmations, instructions, or rapport-building moves. For example, a sentence like "Can you please tell me when the pain started?" was tagged as both a question and an elicitation strategy. This approach allowed for a nuanced breakdown of speech functions common in diagnostic dialogues.

Then the cleaned corpus was uploaded to Sketch Engine, a corpus query analysis software, and, as Figure 2 indicates, the following features were examined:

Word Frequency Lists: To identify commonly used medical terms and discourse markers.

Keyword Analysis: To highlight statistically significant words in comparison with general English corpora.

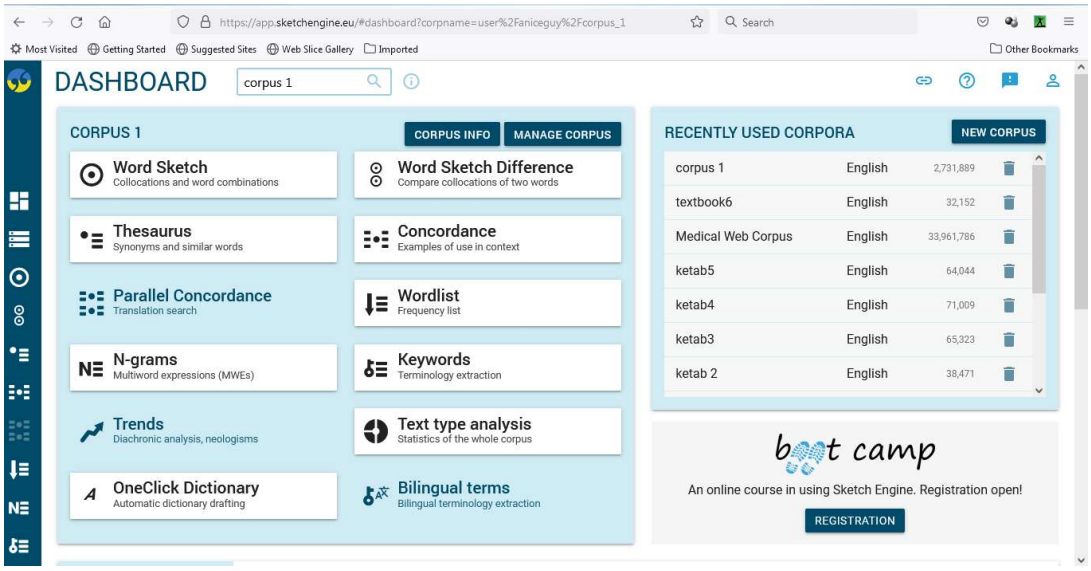
Collocations and Concordances: To observe how specific medical terms are contextually embedded in discourse.

As Figure 2 demonstrates, quantitative analysis was performed using Sketch Engine, focusing on frequency, collocation, and dispersion measures relevant to discourse features. As part of the analysis, the study focused on multi-word expressions like '*you can see*' '*it's important to*' and '*let's talk*.' These expressions are crucial for functional fluency in clinical interactions. Part-of-Speech Tagging and Lemmatization were also examined to analyze grammatical patterns across discourse types. Otter.ai transcription logs showed over 28408 speaker turns across the entire corpus. Speaker-tagging was particularly useful in identifying differences between institutional roles (e.g., doctor, nurse, patient), allowing for a comparative analysis of directive and interrogative structures. For example, doctors used directives like

'*Please lie down*' while patients predominantly employed hedging (e.g., '*I think it started yesterday*').

Figure 2

A Snapshot of the Dashboard Page of the Sketch Engine Website



In the study, the corpus was also split into subsets according to cultural origin (e.g., North American, UK, India, Middle East). In each subset, the utterances were categorized according to their pragmatic functions, and then discourse features were quantified and compared across regions.

To establish inter-rater reliability, two trained annotators independently tagged a subset of the data. Discrepancies were resolved through consensus discussions, and Cohen's Kappa coefficient of 0.79 was calculated, which ensured coding consistency. This step added methodological rigor and minimized subjective bias in the qualitative annotations.

4. Results

The analysis of the YouTube-based medical English corpus yielded several insights into the linguistic features and discourse strategies commonly employed in authentic medical interactions. The results from the Sketch Engine analysis are presented in the following section.

The frequency analysis yielded over 4,000 unique medical terms, with approximately 530 appearing with high frequency across contexts. Table 2 shows high-frequency medical terms.

Table 2

High-Frequency Medical Terms in the Medical Corpus

Medical Term	Frequency
blood	8980
patient	7167
cells	6525
heart	4328
pressure	2529
glucose	2194
insulin	2188
disease	2120
brain	1959
muscle	1920

As seen in Table 2, the most frequent domain-specific terms in the corpus include '*blood*', '*patient*', '*cells*' and '*heart*'. These high-frequency items signal their pedagogical importance in training learners to engage with essential diagnostic and procedural language. Discourse markers like '*and'*, '*so'*, '*or'*, '*as'* and '*but*' were also frequently used to manage topic transitions, frame explanations, and maintain patient engagement. Table 3 demonstrates some of these frequent discourse markers.

As Table 3 indicates, discourse markers such as 'and,' 'so,' and 'or' were pervasive throughout the corpus. These expressions help organize speech, signal transitions, and maintain engagement in clinical explanations. Additionally, as Table 4 shows, Sketch Engine collocation lists revealed frequent verb-noun pairs such as "administer medication," "recommend treatment," and "monitor progress." These patterns can be developed into classroom activities for phrase-building and oral rehearsal.

According to Table 4, collocational patterns such as 'blood pressure,' 'immune system,' and 'blood flow' reflect meaningful phrase combinations used in medical contexts. In Addition, passive constructions (e.g., "*the patient was examined*" and "*medication was administered*")

were prevalent in procedural narratives. Present simple and past simple tenses were dominant, reflecting the need for precision and clarity in ongoing and completed actions. Table 5 indicates key grammar patterns in the medical corpus.

Table 3

Frequent Discourse Markers and Pragmatic Expressions

Discourse Marker	Frequency
and	87169
so	30448
or	15053
as	13972
but	13121
if	12909
now	12382
then	9335
because	8759
also	6345

As Table 5 shows, the grammar of medical discourse frequently relies on passive voice, modal verbs, and present progressive forms. These grammatical structures are essential in describing procedures and giving instructions. Also, as part of the analysis, the current study examined interactional dynamics. Dialogic exchanges, especially in doctor-patient conversations, showcased politeness strategies, question-answer sequences, and empathic communication.

According to Table 6, interactional expressions like '*now let's*,' '*it is important to*' and '*we're going to*' illustrate how clinicians maintain rapport and clarity in discourse. They are key to teaching pragmatics in patient interactions. Comparison of the medical corpora to the British National Corpus (BNC) and other general English corpora like Medical Web Corpus (MWC) in terms of the frequency of field-related keywords was also conducted. Table 7 demonstrates the result of this analysis.

Table 4

Structured Medical Collocations (Lexical Only)

Collocation	Frequency
Blood pressure	880
Immune system	615
Blood flow	543
Heart rate	452
Heart failure	376
Blood vessels	366
Heart disease	205

Table 5

Key Grammar Patterns in Medical Corpus

Grammar Pattern	Frequency
Passive Voice	13562
Present simple passive	10589
Past simple passive	1458
Present perfect passive	445
Past perfect passive	41
Future passive	761
Present continuous passive	243
Past continuous passive	25
Tense	
Simple present	98805
Simple past	11268
Present continuous	10758
Past continuous	1592
Present perfect	1783
Modal Verbs	24188
Can	16261
May	3258

Table 6

Frequent Interactional Phrases in Medical Corpus

Interactional Phrase	Frequency
We're going to	1829
Let me	1148
You can see	1089
I want to	730
Now let's	513
You need to	468
Let's talk	440
It's important to	118
How do you	126
Would you	88

Table 7

The Most Frequently Used Field Specific Vocabulary Comparison Between the Medical Corpus and Reference Corpus

YouTube corpus			Medical Web corpus		
Words	AF*	FinM*	Words	AF	FinM
blood	8980	3223	patient	96419	2292
patient	7167	1519	case	71459	1699
cells	6525	2342	disease	65828	1565
heart	4328	1553	time	56006	1331
pressure	2529	907	treatment	53742	1277
glucose	2194	787	year	42728	1016
insulin	2188	785	day	40561	964
disease	2120	760	cell	37240	885
brain	1959	703	child	36595	780
muscle	1920	387	blood	36381	865

Note. *AF = Absolute Frequency; *FinM = Frequency in Millions. AF indicates the raw count of word occurrences within the corpus, while FinM normalizes these counts per million words to enable cross-corpus comparison.

According to Table 7, the results that are reported in absolute frequency (AF) and frequency in millions (FinM) show significantly higher relative usage of terms like *cell*, *blood*, *patient*, *body*, and *heart*, confirming the domain-specific concentration of the dataset (only 10 frequent words have been reported for two corpora due to space limitations in the manuscript). These words reflect core concepts in diagnosis, patient care, and treatment, making the corpus highly valuable for EMP instruction.

Interestingly, the corpus revealed noticeable differences in discourse style based on cultural origin. As Table 8 shows, videos from North American institutions emphasized shared decision-making and empathy, whereas some UK or Indian content showed more doctor-led instruction.

Table 8

Cultural Variation in Medical Discourse Styles

Region	Discourse Style	Empathy Level	Common Phrases
North America (e.g., USA, Canada)	Shared decision-making, high empathy, inclusive language	High	Often uses 'we,' 'let's,' 'how do you feel?'
United Kingdom	Doctor-led discussion with polite formality	Moderate	Often uses 'you need to,' 'shall we,' 'let's proceed'
India	Directive style, authoritative tone, patient deference	Low	Often uses 'you must,' 'take this,' less hedging or questioning
Middle East	Polite yet directive, culturally respectful	Moderate	Tends toward formal medical authority with patient politeness

As Table 8 indicates, North America (e.g., USA, Canada) shared decision-making, high empathy, inclusive language like '*we*', 'let's' and '*how do you feel*?' whereas the United Kingdom demonstrated doctor-led discussion with polite formality like '*you need to*', '*shall we*' and '*let's proceed'*. India characterized a directive style, authoritative tone, patient deference, and al level of empathy such as '*you must take this*' less hedging or questioning.

Middle East featured a polite yet directive, culturally respectful, and a moderate level of empathy with a tendency toward formal medical authority and polite address.

5. Discussion

The findings of this YouTube-based corpus study provide compelling evidence that discourse-level features are central to professional communication in medical settings and must be foregrounded in EMP instruction. This aligns with a growing body of research in ESP (Basturkmen, 2010; Hyland, 2004; Swales, 1990) which calls for a shift from traditional form-focused instruction to approaches that prioritize genre awareness, interactional competence, and contextualized language use.

The corpus, drawn from authentic YouTube videos featuring medical consultations, health explanations, and patient education, revealed distinct discourse patterns that recur across different speakers, specialties, and cultural contexts. These patterns include high-frequency multiword units, noun–noun and verb–noun collocations, discourse markers, modal verbs, passive constructions, and interactional phrases. Such features are not random but rather reflect the conventionalized ways in which medical professionals structure talk, manage interpersonal relationships, and navigate complex clinical tasks.

A prominent feature in the corpus is the use of formulaic language, including fixed expressions like "*take a deep breath*" or "*on a scale from 1 to 10*". These are examples of what Biber et al. (2004) termed "lexical bundles" — multiword sequences that perform crucial discourse functions such as introducing procedures, giving reassurance, or eliciting diagnostic information. Flowerdew (2004) emphasizes that such patterned language is essential for fluency and communicative efficiency in specialized domains, especially where time, clarity, and trust are critical, as in healthcare.

The frequency of noun–noun collocations such as 'heart rate,' 'blood pressure,' 'glucose level,' and 'muscle tension' is consistent with Cheng's (2012) findings on the lexical density of professional genres. These collocations not only convey complex concepts succinctly but also function as disciplinary identifiers that signal expertise. Their correct usage

contributes significantly to a speaker's perceived credibility, a factor of particular relevance in clinical communication where patient trust is foundational.

The study also found a notable density of passive constructions, especially in procedural contexts. Passive voice structures such as "*you will be given*…," "*the test is performed*…," and "*medication is prescribed*…" help maintain a focus on procedures and outcomes rather than agents. While often discouraged in general English instruction, passive constructions serve pragmatic functions in medical discourse: they depersonalize actions, mitigate directness, and shift the emphasis to patient care processes (Gotti, 2008). These findings confirm Bhatia's (1993) observations about professional genres relying heavily on passive voice to conform to institutional norms of objectivity and detachment.

In terms of grammar and modality, the corpus revealed significant use of modal verbs — particularly *can, should, must,* and *may* — which are used to express ability, obligation, suggestion, and possibility. This aligns with Paltridge's (2006) argument that modality is central to hedging and managing epistemic stance in academic and professional discourse. For example, expressions such as "*you can take this medication*" or "*you should avoid fatty foods*" offer guidance while allowing room for patient autonomy. Such use of modals reflects a balanced communication style that combines clinical authority with patient-centeredness.

One of the more nuanced aspects of the data is the high occurrence of interactional phrases and discourse markers. These include expressions like "*we're going to do*", "*you might need surgery*" and "*How are you feeling today*?" According to Schiffrin (1987) and McCarthy (1998), such markers serve as meta-discourse tools—they help manage the flow of information, guide listeners through complex talk, and establish interpersonal rapport. In the medical context, they also function as discourse regulators, helping professionals manage time, clarify expectations, and signal transitions in interaction.

Moreover, the frequent use of these interactional elements reflects the dialogic nature of clinical encounters. Medical consultations are not unidirectional but involve a co-construction of meaning between doctor and patient. This resonates with Hyland's (2004) framework of stance and engagement, where writers and speakers use language to both assert their authority and invite participation. In face-to-face settings, engagement strategies are

crucial in negotiating compliance, checking understanding, and responding empathetically to patient concerns.

An important finding with pedagogical implications is the strong presence of cultural variation in medical discourse across different videos. For instance, clinicians from Western contexts were more likely to use mitigated directives, indirect requests, and hedging strategies (e.g., "*I would recommend…*"), whereas those from South Asian or Middle Eastern settings exhibited more direct and authoritative styles (e.g., "*You must take this now*"). This observation supports the work of Scollon and Scollon (2001) on intercultural communication in professional discourse, suggesting that EMP instruction should include cross-cultural pragmatics as a core component.

The use of YouTube as a data source also deserves attention. As Boulton (2012) and Gavioli (2005) argue, video corpora offer advantages that written texts or simulated dialogues cannot match. They present authentic language in action, complete with intonation, gestures, turn-taking behavior, and spontaneous repairs — all of which are integral to pragmatic competence. Moreover, YouTube provides a globally diverse sample of English speakers in real-life settings, offering exposure to World Englishes and intercultural variation. This is particularly relevant for international medical students who will likely engage with a linguistically and culturally diverse patient population.

6. Conclusion and Recommendations

This study has demonstrated that YouTube-based corpora can serve as a powerful tool in developing discourse competence among medical English learners. By analyzing real-world examples of medical communication, students can gain exposure to the pragmatic, grammatical, and rhetorical patterns that characterize professional medical discourse. The consistent appearance of domain-specific terminology, structured interactional patterns, and discourse markers suggests that authentic spoken language provides clearer and more accessible learning targets than traditional, scripted resources. The incorporation of these findings into EMP instruction can lead to improved outcomes for learners aiming to function effectively in clinical or academic settings.

From a curricular standpoint, the results advocate for a more discourse-centered approach to ESP instruction. Rather than isolated vocabulary lists or grammar drills, language programs should prioritize exposure to authentic interactional routines. Corpus-informed syllabi can equip learners with practical communicative tools derived from actual professional discourse. In other words, the findings emphasize the need for ESP syllabi to shift from traditional grammar-focused instruction to discourse-oriented strategies.

Although discourse markers such as *and, so, or, as*, and *but* are common in general spoken English, their functions within the medical corpus reveal distinctive genre-specific uses. In everyday conversation, such markers typically serve interpersonal or casual cohesive purposes (McCarthy, 1998; Schiffrin, 1987). In contrast, in professional medical discourse, they operate as organizational and instructional devices: *and* frequently sequences diagnostic or procedural steps ("*take a deep breath and hold it*"), *so* introduces explanations or inferences ("*so, this indicates inflammation*"), and *but* mitigates contrastive statements while maintaining empathy ("*your cholesterol is high, but it can be managed*"). These pragmatic patterns indicate that, while formally similar, discourse markers in medical interaction contribute to clarity, reassurance, and structured communication — functions that are less salient in informal genres. The high frequency of passive constructions, modal verbs, and progressive aspects in the corpus reveals that grammar in real-world medical contexts serves specific communicative functions such as emphasizing patient conditions, expressing uncertainty, or describing ongoing processes rather than merely following prescriptive rules. Furthermore, the prevalence of collocations and fixed lexical patterns underscores the need to teach grammar as part of larger, meaningful language chunks rather than in isolation. Interactional phrases like "*let me explain*" or "*can you see*" which are essential for guiding and engaging listeners, highlight the importance of incorporating pragmatic competence and interpersonal strategies into instruction. Most significantly, the clear cultural variation in discourse styles—ranging from empathetic and inclusive language in North America to more directive patterns in other regions—demonstrates that learners must develop sensitivity to context, tone, and audience. These findings justify the move toward discourse-based teaching methods that integrate grammar with real-life use, communicative purpose, and intercultural awareness—ultimately equipping learners with the skills necessary for effective, patient-centered communication in

global medical settings. Future research might explore how repeated exposure to discourse patterns enhances long-term retention and clinical communication performance.

This study emphasizes the necessity for long-term strategies in corpus integration, such as aligning corpus-driven materials with established pedagogical frameworks like the Common European Framework of Reference (CEFR). While CEFR levels were not directly analyzed in this research, the framework offers a standardized point of reference for calibrating corpus-based medical English instruction to internationally recognized proficiency outcomes. Educational stakeholders should invest in professional training that equips ESP instructors with corpus analysis skills, ensuring sustainability in corpus-enhanced teaching.

Limitations of the current study include the reliance on publicly available YouTube content, which may not fully represent private or emergency clinical discourse. Nonetheless, this source offers a wide spectrum of high-quality data with global accents, settings, and practices, making it broadly applicable for international medical English instruction. Future studies could expand upon the present findings by incorporating quantitative learner performance data to measure the direct impact of corpus-based instruction on communicative outcomes in medical English. Moreover, the integration of multimodal corpora— including gesture, prosody, and visual cues —would provide a more comprehensive understanding of how meaning is constructed in clinical interaction. Longitudinal studies are also needed to examine how sustained exposure to authentic discourse patterns influences learners' pragmatic competence and patient-centered communication over time.

Conflict of interest

The author(s) certify/certifies that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in the present research paper.

References

Adolphs, S., & Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimodal.* Routledge.

Baker, P. (2006). *Using corpora in discourse analysis.* Continuum.

Baldry, A., & Thibault, P. J. (2006). *Multimodal transcription and text analysis: A multimedia toolkit and coursebook*. Equinox.

Bardovi-Harlig, K., & Hartford, B. S. (1993). Learning the rules of academic talk: A longitudinal study of pragmatic change. *Studies in Second Language Acquisition, 15*(3), 279–304.

Basturkmen, H. (2010). *Developing courses in English for specific purposes.* Macmillan.

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings.* Routledge.

Biber, D., Conrad, S., & Reppen, R. (1998*). Corpus linguistics: Investigating language structure and use.* Cambridge University Press.

Boulton, A. (2017). Corpora in language teaching and learning: Learner engagement, motivation, and autonomy. *Language Learning Journal, 45*(5), 582–597. https://doi.org/10.1080/09571736.2017.1312491

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47.

Candlin, C. N., & Candlin, S. (2002). Discourse, expertise and the management of risk in health care settings. *Research on Language and Social Interaction, 35*(2), 115–137.

Chan, K. L. (2018). Communicative competence in the medical field: A corpus-based analysis. *International Journal of English Linguistics, 8*(3), 46–58.

Chen, Y. -H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing: Applications of corpus tools in the study of disciplinary variation. *Journal of English for Academic Purposes, 9*(3), 183–197. https://doi.org/10.1016/j.jeap.2010.02.001

Csomay, E., & Petrovic, M. M. (2012). "Yes, your Honor!": Courtroom discourse and the corpus-informed classroom. *English for Specific Purposes, 31*(2), 111–121.

Dooly, M. A. D., & O'Dowd, R. (Eds.). (2018). *In this together: Teachers' experiences with transnational, telecollaborative language learning projects.* Peter Lang Ltd.

Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional language. In U. Connor & T. A. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 11–33). John Benjamins Publishing Company. https://doi.org/10.1075/scl.16.02flo

Flowerdew, L., & Wang, S. (2015). Teaching discourse features of academic lectures through multimodal corpora. *English for Specific Purposes, 37,* 59–72. https://doi.org/10.1016/j.esp.2014.06.001

Gavioli, L. (2005). *Exploring corpora for ESP learning*. John Benjamins Publishing Company.

Godwin-Jones, R. (2018). Emerging technologies: Using YouTube for language learning. *Language Learning & Technology, 22*(1), 1–15.

Gotti, M. (2008). Investigating specialized discourse. In F. I. Gómez & C. A. Räisänen (Eds.), *ESP in European higher education: Integrating language and content* (Vol. 4) (pp. 17–34). John Benjamins Publishing.

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Hafner, C. A., & Miller, L. (2011). Fostering learner autonomy in English for Science: A collaborative digital video project in a technological learning environment. *Language Learning & Technology, 15*(3), 68–86.

Hutchinson, T., & Waters, A. (1987). *English for specific purposes: A learning-centred approach.* Cambridge University Press.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. University of Michigan Press.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing.* Continuum.

Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41–62. https://doi.org/10.1111/j.1473-4192.2008.00179.x

Hyland, K. (2019). Second language writing and ESP: The challenges of engagement. *Journal of English for Academic Purposes, 38*, 75–85. https://doi.org/10.1016/j.jeap.2019.01.004

Jiang, Y., & Peng, J. E. (2025). Exploring the relationships between learners' engagement, autonomy, and academic performance in an English language MOOC. *Computer Assisted Language Learning, 38*(1-2), 71-96.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

McCarthy, M. (1998). *Spoken language and applied linguistics.* Cambridge University Press.

Nguyen, T. M., & Boers, F. (2021). Using authentic audiovisual input to foster conversational strategy awareness in EFL learners. *System, 98,* 102462.

Otter.ai. (2023). *Otter.ai* [AI-powered transcription software]. https://otter.ai

Paltridge, B. (2006). *Discourse analysis: An introduction.* Continuum.

Roberts, C., & Sarangi, S. (2005). Theme-oriented discourse analysis of medical encounters. *Medical Education, 39*(6), 632–640. https://doi.org/10.1111/j.1365-2929.2005.02171.x

Sarangi, S., & Roberts, C. (Eds.). (1999). *Talk, work and institutional order: Discourse in medical, mediation and management settings*. Mouton de Gruyter. https://doi.org/10.1515/9783110208375

Schiffrin, D. (1987). *Discourse markers.* Cambridge University Press.

Scollon, R., & Scollon, S. W. (2001). *Intercultural communication: A discourse approach* (2nd ed.). Blackwell Publishers.

Sketch Engine. (2023). *Sketch Engine* [Corpus linguistics software]. https://www.sketchengine.eu

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* John Benjamins.

Tribble, C. (2015). Writing academic English: A corpus-based perspective. *Language Teaching, 48*(3), 354–368. https://doi.org/10.1017/S0261444815000155

Tseng, F. P. (2013). Medical English corpus analysis: A case study of doctor–patient communication. *Asian ESP Journal, 9*(2), 90–117.

Wang, W. (2010). The discourse of medical consultations in China: A genre-based study. *Discourse & Communication, 4*(2), 203–225.

YouTube. (2023). *YouTube* [Video-sharing platform]. https://www.youtube.com