

A CORPUS-BASED ANALYSIS OF DISTRIBUTIONAL AND FUNCTIONAL DIFFERENCES IN DISCOURSE MARKERS ACROSS SPOKEN AND ACADEMIC ENGLISH

RESEARCH ARTICLE

Dr. Ferdi ÇELİK

Ondokuz Mayıs Üniversitesi

ferdicelik99@gmail.com

ORCID: 0000-0001-8272-4720

DOI: [10.5281/zenodo.18158228](https://doi.org/10.5281/zenodo.18158228)

Gönderim Tarihi: 09.10.2025 Kabul Tarihi: 22.12.2025

Alıntı: ÇELİK, F. (2025). A Corpus-Based Analysis of Distributional and Functional Differences in Discourse Markers across Spoken and Academic English. *Türkeri Dil ve Edebiyat Dergisi (TürkeriDE)*, 2, 13-22.

ABSTRACT: This corpus-based study examines the extent to which selected discourse markers (*well*, *you know*, *actually*, *I mean*, *anyway*, *in fact*, *indeed*) differ in distribution and functional realization across spoken and academic written English in the Corpus of Contemporary American English. Using the spoken (126,135,576 words) and academic (119,790,456 words) sections in the corpus, the analysis extracted raw token counts and normalized frequencies per million words, and computed spoken-to-academic ratios as a register-sensitivity index. To address polyfunctionality, a targeted concordance validation was conducted for *well* and *actually*, sampling 10 lines per register (20 per marker) and coding each instance as discourse-marker/discourse-marker-like versus non-discourse marker, with an indicative adjusted discourse marker rate derived from observed proportions. Results show strong register differentiation: *you know*, and *I mean* display the largest spoken skew, *anyway*, is also spoken-preferential, while *well* and *actually* show moderate spoken skews in raw counts. *In fact* exhibits only a modest spoken preference, whereas *indeed* is the sole marker favoring academic prose. Functional validation indicates that raw frequencies substantially overestimate DM use for polyfunctional items, especially *actually*, whose discourse-pragmatic uses are minority instances in both registers. Pedagogically, the findings support corpus-informed instruction that treats discourse markers as register-bound resources and incorporates concordance-based activities to build learners' pragmatic and register competence.

Keywords: discourse markers; register variation; corpus-based analysis; COCA; spoken and academic English; functional realization; EFL

Konuşma Dili İngilizcesi ile Akademik İngilizce Arasında Söylem Belirteçlerindeki Dağılımsal ve İşlevsel Farklılıkların Derlem Temelli Bir Analizi

ÖZ: Bu derlem temelli çalışma, seçilmiş söylem belirteçlerinin (*well*, *you know*, *actually*, *I mean*, *anyway*, *in fact*, *indeed*) konuşma dili İngilizcesi ile akademik yazılı İngilizce arasında dağılım ve işlevsel gerçekleşim bakımından ne ölçüde farklılaştığını Corpus of Contemporary American English (COCA) verileri üzerinden incelemektedir. Çalışmada, derlemin konuşma dili (126.135.576 sözcük) ve akademik (119.790.456 sözcük) bölümleri kullanılarak ham sözcük sıklıkları çıkarılmış, milyon sözcük başına normalize edilmiş frekanslar hesaplanmış ve konuşma dili-akademik oranları, tür (register) duyarlılığını gösteren bir indeks olarak değerlendirilmiştir. Çok işlevlilik (polyfunctionality) sorununu ele almak amacıyla *well* ve *actually* için hedefli bir dizin (concordance) doğrulaması yapılmış; her bir türden 10 satır (belirteç başına toplam 20 örnek) örneklenmiş ve her kullanım söylem belirteci/söylem belirteci benzeri ya da söylem belirteci olmayan olarak kodlanmıştır. Gözlenen oranlara dayanarak gösterge niteliğinde düzeltilmiş bir söylem belirteci kullanım oranı türetilmiştir. Bulgular, güçlü bir tür ayrışmasına işaret etmektedir: *you know* ve *I mean* konuşma dilinde en belirgin yoğunlaşmayı gösterirken, *anyway* de konuşma dilini tercih eden bir kullanım sergilemektedir. *Well* ve *actually* ham sayılarda orta düzeyde konuşma dili eğilimi göstermektedir. *In fact* yalnızca sınırlı bir konuşma dili tercihinin sahipken, *indeed* akademik yazıyı tercih eden tek belirteçtir. İşlevsel doğrulama, özellikle *actually* için, ham frekansların söylem belirteci kullanımını önemli ölçüde fazla tahmin ettiğini; bu belirtecin söylem-edimsel kullanımının her iki türde de azınlıkta kaldığını göstermektedir. Eğitsel açıdan, bulgular söylem belirteçlerinin türe bağlı kaynaklar olarak ele alınmasını ve öğrenenlerin edimsel ve türel yeterliklerini geliştirmek amacıyla dizin temelli etkinliklerin öğretime dâhil edilmesini desteklemektedir.

Keywords: söylem belirteçleri; tür değişkenliği; derlem temelli analiz; COCA; konuşma dili ve akademik İngilizce; işlevsel gerçekleşim; EFL

Introduction

Discourse markers (DMs) such as *well*, *you know*, *actually*, *I mean*, *in fact*, and *indeed* are essential elements of spoken and written communication. They function as pragmatic cues that organize discourse, indicate speaker stance, and facilitate coherence between utterances (Crible, 2017; Cuenca & Crible, 2019). In spoken discourse, DMs often manage turn-taking, repair, and interactional alignment, while in academic writing, they signal logical relationships, mark evidentiality, and frame argumentation

(Zhou & Xia, 2024). Despite their communicative importance, DMs remain underrepresented in pedagogical and applied linguistic contexts, particularly in English language teaching (ELT), where register variation is seldom explicitly addressed (Le Foll, 2024; Aşık, 2015).

The Corpus of Contemporary American English (COCA) offers an empirical basis to the study of register variation in the use of DM. The systematic comparison of the differences in the distribution and functioning of DMs is possible due to its balanced coverage of spoken and academic written registers. These analyses play a vital role in the gap between theoretical linguistics and pedagogy, as it is shown that EFL learners do not use or misuse DMs because of insufficient input and teaching (Vickov & Jakupcevic, 2017; Neary-Sundquist, 2014).

The knowledge of the variation of DMs between registers, therefore, is important in corpus-based linguistic studies as well as the development of pragmatic competence in ELT. Since learners are exposed to the English language in more and more contexts, academic writing, or conversational fluency, the knowledge of DM use in real corpora is a potent means of teaching contextualized communication (Cushing, 2017).

Literature Review

Discourse markers are typically considered to be procedural features indicating relationships between discourse segments without any change in propositional content (Crible, 2017; Crible & Cuenca, 2019). Pragmatic theories place DMs in the textual, interpersonal, and cognitive space (Crible & Pascual, 2020). They are in charge of controlling coherence, repairing, and directing listeners to interpret (Shustova & Tsarenko, 2019). Cognitively linguistically, DMs are formed in a process of grammaticalization and constructionalization, in which lexical items become discourse-organizing over time by repeated contextualization (Traugott, 2021). This explains the polyfunctionality of markers like “*well*” and “*actually*”, which can signal stance, correction, or topic shifts depending on context (Crible & Blackwell, 2020). DMs also reflect speakers’ probabilistic competence, which can be defined as the ability to select forms appropriate to context, audience, and modality (Engel et al., 2022). In this context, DMs are dynamic structures that are influenced by social interaction, communicative objectives, and contextual limitations.

Register theory offers an alternative point of view. According to the multidimensional model by Biber (2002), linguistic variation is determined as functional adjustment to situational parameters, including mode, tenor, and field. Spoken registers are more concerned with participation and mutual understanding, whereas written academic registers are more concerned with the density of the information and its abstraction (Biber et al., 2002; Szmrecsanyi, 2019). Systemic Functional Linguistics (SFL) DMs are interpersonal metaphors of modality, which can be used to show commitment, evaluation, and rhetorical stance (Cuenca, 2022). For instance, *in fact*, and *indeed* strengthen claims epistemically, while *well* and *you know* mitigate or contextualize interpersonal intent. Corpus-based research in the recent past has revealed that there is a systematic difference in modality and stance expressions in English across registers, which confirms that DMs are distributed functionally (Zhou & Xia, 2024). The fact that such register differences are almost universal is further supported by cross-linguistic corpus studies, which show that orality is associated with the frequent, polyfunctional use of DM, whereas formal writing limits the scope and functionality of such distinctions (Schnell & Schiborr, 2022; Shirtz, 2024).

The literature of corpus has drawn definite differences in DM frequency and role in registers. Research on COCA, the British National Corpus, and DisFrEn has established that DMs are one of the most common aspects that differ between spoken language and written prose (Jucker, 2021; Crible, 2017). Spoken English is rich in interactive markers like “*you know*”, “*I mean*”, and “*well*”, which facilitate real-time discourse management, whereas academic texts rely on cohesive devices such as “*in fact*”, “*indeed*”, and “*however*” for logical progression (Biber et al., 2002; Zhou & Xia, 2024). A comparative study of English and French by Crible (2017) showed that DMs are often co-occurring with disfluencies, pauses, and repairs and serve as cognitive scaffolds in the speech production process. In the same vein, Cuenca and Crible (2019) discovered that patterns of co-occurrence of DMs are recurrent, which proves that such markers as *well* and *so* are functional within discourse sequences.

The difference in discourse markers among the varieties of English and other languages highlights the sociolinguistic embeddedness of these discourse markers. A comparative study of Zimbabwean and

British English revealed significant pragmatic differences in the use of *so* and *well*, reflecting local discourse norms (Chapwanya & Nel, 2024). Apostol et al. (2025) confirmed that multiword DMs such as *you know* and *in fact* vary across languages, correlating with genre and register. Typological corpus studies extend these insights, suggesting that DMs universally mark cohesion and narrative structure, though their lexical realizations differ (Schnell & Schiborr, 2022). In Catalan, for instance, shifts in the norms of DMs like *de fet* (“in fact”) illustrate sociolinguistic change influenced by contact and genre (Cuenca, 2022). All these findings confirm that DM use is a manifestation of cognitive universals, as well as culture-specific patterns of communication.

In applied settings, DMs are key indicators of communicative competence. Studies on learner corpora show that L2 speakers’ DM usage increases in range and appropriateness with proficiency (Neary-Sundquist, 2014). Aşık (2015) demonstrated that explicit instruction in DM use enhances pragmatic awareness and fluency among Turkish EFL learners. Similarly, in classroom discourse, non-native teachers often underutilize interactive DMs, relying on a restricted set of connectors (*so*, *ok*, *and*) for discourse organization (Vickov & Jakupčević, 2017). This is in contrast to the native speaker talk, where DMs have a broad range of interpersonal and cohesive roles. The latest studies support the idea of corpus-based teaching that uses real-life register information to educate DMs in context (Cushing, 2017; Le Foll, 2021). The interaction of the learners with the corpus examples promotes pragmatic sensitivity and assists in overcoming the distance between formal learning and spontaneous communication.

DMs are mainly employed in academic registers to organize the argumentation and express epistemic position. Studies of student writing reveal consistent underuse of metadiscourse markers like *indeed*, *therefore*, and *in fact* (Chung et al., 2023). Barbara et al. (2024) discovered that beginners are not aware of the encoding of certainty and politeness by stance markers, which leads to excessive assertiveness or incoherent academic writing. Zhou and Xia (2024) demonstrated that COCA’s academic subcorpus displays a high density of nominalized and impersonal modality (*the possibility that, it is likely that*), distinguishing it from spoken registers that rely on personal expressions like *I think* or *you know*. DMs are also crucial to discourse in genres. In theater translation, Tarantini and Benatta (2021) discovered that DMs maintain naturalism and rhythm in a conversation, which supports their role in the natural interaction. Furko (2017) demonstrated in the political context that DMs have manipulative rhetorical purposes, positioning arguments and directing the interpretation of the audience. Together, these studies emphasize that DMs are versatile discourse tools that adapt to communicative purpose and audience expectations.

The Gap and the Present Study

Although extensive studies have been carried out on DMs in languages and genres, there is still no systematic, functionally tested register comparison of DMs in COCA. Earlier research tends to focus on individual markers, small corpora, or even on particular genres without considering large-scale register differences between spontaneous speech and formal academic writing. Furthermore, there are not many studies that combine quantitative frequency analysis with qualitative functional validation- a method that is essential in the differentiation of discourse-pragmatic and lexical uses of polyfunctional items like *well* and *actually*. Moreover, even though corpus linguistics has proven to be pedagogically viable, its use in ELT materials and pragmatic teaching is scarce (Le Foll, 2021). This empirical-practical gap will help us to improve our knowledge about register competence and guide information-based language teaching. Therefore, this study investigates to what extent selected discourse markers (“*well*, *you know*, *actually*, *I mean*, *anyway*, *in fact*, *indeed*”) differ in their distribution and functional realization between the spoken and academic written registers in COCA. Specifically, it aims to:

1. Quantify DM frequency and register ratio differences between spoken and academic English;
2. Functionally validate polysemous DMs (*well*, *actually*) through concordance sampling;
3. Interpret findings pedagogically to inform ELT practices related to register awareness and discourse competence.

By combining quantitative corpus methods with qualitative pragmatic validation, the study provides an empirically grounded model for understanding DM variation across registers and contributes to corpus-informed ELT pedagogy. Therefore, the study is guided by the following research question:

RQ. Is there a difference in the use of markers (e.g., well, you know, actually) between spoken and written registers in COCA, and how should these differences be addressed in ELT?

Methods

Corpus and register design

The study used the Corpus of Contemporary American English (COCA) as the sole data source, with analyses restricted to the 1990–2019 span as represented in the COCA interface. The register comparison was operationalized as a contrast between the SPOKEN and ACADEMIC sections, because this pair most directly instantiates the spoken–written distinction that is central to English language teaching decisions about discourse markers. The section word counts reported by the COCA “compare sections” output were 126,135,576 words for SPOKEN and 119,790,456 words for ACADEMIC, and these values were treated as the denominators for normalized frequency calculations.

Target markers and query specification

Seven high-frequency candidates were selected to represent discourse marking and related pragmatic work often discussed in ELT materials and applied linguistics: you know, I mean, well, actually, anyway, in fact, and indeed. The queries were entered as literal word/phrase searches in COCA (e.g., “YOU KNOW”, “I MEAN”, “IN FACT”), thereby targeting the orthographic realizations most likely to be visible to teachers and learners. Because some items are polyfunctional (most notably well and actually), the study adopted a two-stage logic: (a) establish register distribution through normalized frequencies and (b) validate the DM function where lexical uses could substantially inflate counts.

Frequency extraction and normalization

For each marker, COCA’s section-comparison output was used to obtain (i) raw token counts per section and (ii) normalized frequencies per million words (pmw) for SPOKEN and ACADEMIC. A spoken-to-academic ratio was computed as $PM(SPOKEN)/PM(ACADEMIC)$. Ratios substantially greater than 1 were interpreted as spoken-leaning; ratios substantially less than one were interpreted as academic-leaning. Given the study’s urgent, rapid-conduct design, the ratio was used as the primary effect-size index, and statistical significance testing was not treated as necessary for inference, because the corpus sizes are large and the observed contrasts for several markers are orders of magnitude.

Concordance sampling and functional validation

To prevent misclassification for markers with strong lexical/polyfunctional behavior, a targeted concordance (KWIC) validation was conducted for well and actually. For each of these two items, 10 KWIC lines were extracted from SPOKEN and 10 from ACADEMIC (20 per marker) and coded for whether the token functioned as a discourse marker rather than a lexical item. The coding scheme was intentionally minimal and focused on a single decision boundary:

1. **DM / DM-like:** tokens used for turn initiation, stance framing, correction, repair, topic management, or parenthetical pragmatic orientation (e.g., “Well, ...”; “actually, ...” in a corrective parenthetical).
2. **Non-DM:** tokens functioning propositionally or lexically (e.g., well tempered; do well; actually implemented; actually work), including fixed academic sequences such as as well as.

Because a portion of ACADEMIC data contains embedded spoken genres (e.g., interview transcripts, classroom talk excerpts), DM-like uses in academic KWIC lines were additionally interpreted with genre sensitivity: a DM-like token occurring inside quoted speech was treated as evidence about the presence of conversational discourse within academic texts rather than as evidence that the marker is typical of authorial academic prose.

The KWIC-based proportions were used to derive an indicative “adjusted DM frequency” for well and actually by multiplying pmw by the DM proportion observed in the KWIC sample. This adjustment was

not treated as a definitive estimate; rather, it served to make transparent how far raw frequency might overstate discourse-marker use in writing.

Analytical strategy

The analysis proceeded in three steps. First, the register distribution of each marker was established through PMW and ratios. Second, markers were classified along a continuum from spoken-anchored interactional marking to written-anchored epistemic strengthening, with special attention to polyfunctional items whose raw frequencies risk pragmatic misinterpretation. Third, findings were interpreted with respect to the research question's pedagogical clause ("how should these differences be addressed in ELT?") by inferring which markers are appropriate targets for spoken fluency and interactional competence training versus which belong primarily to written argumentation and academic style.

Results

Results are presented in two blocks aligned with the analytic workflow. First, register distributions for seven candidate markers were reported using raw token counts and normalized frequencies per million words (pmw) in the COCA SPOKEN and ACADEMIC sections, together with a spoken-to-academic ratio. Second, because *well* and *actually* are polyfunctional, a targeted concordance-based functional check was conducted for these two items using 10 KWIC lines per register (20 per item), coded as discourse-marker (DM/DM-like) versus non-DM.

Table 1 reports the distributional results. Six of the seven markers occurred at higher pmw in SPOKEN than ACADEMIC, while *indeed* showed the reverse pattern.

Table 1. Register distribution of candidate markers in COCA (SPOKEN vs ACADEMIC)

Marker	SPOKEN tokens	SPOKEN pmw	ACADEMIC tokens	ACADEMIC pmw	Ratio (SPOK/ACAD)
you know	348,539	2,763.2	3,209	26.8	103.1
I mean	130,528	1,034.8	1,227	10.2	101.0
well	363,946	2,885.4	90,012	751.4	3.8
actually	89,657	710.8	16,157	134.9	5.3
anyway	11,195	88.8	1,205	10.1	8.8
in fact	35,930	284.9	18,976	158.4	1.8
indeed	10,450	82.8	22,638	189.0	0.4

Two multiword markers, *you know* and *I mean*, displayed the largest spoken–academic contrasts (ratios =103 and =101, respectively). *Anyway* also patterned as spoken-preferential (ratio =8.8), while *well* and *actually* showed smaller but still notable spoken skews (ratios =3.8 and =5.3). In contrast, *in fact* showed a relatively modest spoken preference (ratio =1.8). *Indeed* was the only marker with higher normalized frequency in ACADEMIC than SPOKEN (ratio =0.4, corresponding to an academic-to-spoken ratio of =2.3).

Because raw counts for *well* and *actually* may include substantial non-discourse uses, KWIC samples were coded for pragmatic function. Table 2 reports DM/DM-like versus non-DM classifications and the resulting DM proportions; an indicative adjusted DM pmw is included by multiplying the raw pmw by the observed DM proportion in the sample.

Table 2. KWIC-based functional validation for polyfunctional markers (*well*, *actually*)

Marker	Register	KWIC lines	DM/DM-like	Non-DM	DM proportion	Raw pmw	Indicative DM pmw	adjusted
<i>well</i>	SPOKEN	10	7	3	0.70	2,885.4	2,019.8	
<i>well</i>	ACA-DEMIC	10	3	7	0.30	751.4	225.4	
<i>actually</i>	SPOKEN	10	1	9	0.10	710.8	71.1	
<i>actually</i>	ACA-DEMIC	10	2	8	0.20	134.9	27.0	

In the SPOKEN KWIC sample for *well*, 7 of 10 instances were coded as DM/DM-like, while 3 were coded as non-DM. In the ACADEMIC KWIC sample for *well*, 3 of 10 instances were coded as DM/DM-like and 7 as non-DM; non-DM instances included additive and lexical realizations (e.g., *as well as*, evaluative lexical uses), while DM/DM-like realizations occurred in segments presented as reported or quoted talk within academic texts.

For *actually*, the SPOKEN KWIC sample contained 1 DM/DM-like instance and 9 non-DM instances, and the ACADEMIC sample contained 2 DM/DM-like instances and 8 non-DM instances. Non-DM instances in both registers primarily involved truth-conditional adverbial modification (e.g., *actually implemented*, *actually work*). DM/DM-like instances in ACADEMIC were observed in quoted or rhetorically marked sentence-initial environments within the sampled lines.

Consequently, the distributional analyses indicated substantial register differentiation across markers, with markedly higher spoken frequencies for *you know* and *I mean*, moderate spoken preference for *anyway*, smaller spoken skews for *well* and *actually*, comparatively limited spoken preference for *in fact*, and academic preference for *indeed*. The functional validation further showed that, for *well* and *actually*, raw frequencies included a substantial share of non-discourse uses in both registers, with DM/DM-like uses constituting a larger proportion of *well* in SPOKEN than in ACADEMIC, and constituting a minority of *actually* in both registers within the examined samples.

Discussion

The findings indicated a high and stable register distinction in the distribution and pragmatic expression of DMs in the spoken and academic written English in the COCA. Six of the seven markers discussed (*you know*, *I mean*, *well*, *actually*, *anyway*, *in fact*) were much more frequently normalized in the spoken register, and did have a reverse, scholarly preference. *You know* and *I mean* were the most markedly different, having a ratio of spoken-to-academic more than 100, and indicating their prototypical connection with informal, interactive discourse.

Functional validation of polyfunctional items (*well* and *actually*) by concordance further explained that raw frequency data exaggerates DM use in writing since most of the instances are lexical or truth-conditional but not discourse-pragmatic. *Well* retained a significant spoken dominance, after qualitative adjustment, and *actually* emerged as discourse-like only marginally in both registers. These results demonstrate that DM distribution is quantitatively and functionally register-sensitive, which is consistent with well-known models of register variation (Biber et al., 2002; Szmrecsanyi, 2019).

The register contrast of the current study supports and expands on previous studies in the field of corpus linguistics that separate the interactional spoken discourse and the informational written discourse. *You know* and *I mean* dramatic frequency skew are similar to those of previous studies that report that these multiword markers are canonical features of involvement, intersubjectivity, and turn-management in speech (Jucker, 2021; Crible & Cuenca, 2019). Their rarity in academic prose reflects the suppression

of interpersonal stance and dialogic engagement in formal registers (Biber et al., 2002; Zhou & Xia, 2024). Markers such as *anyway* and *well* occupy an intermediate functional position: while predominantly spoken, they serve broader textual or transitional roles that occasionally extend into quasi-spoken or quoted segments of academic texts. This continuum-based distribution mirrors Crible's (2017) observation that DMs cluster along pragmatic dimensions rather than forming discrete categories. The moderate ratios for *anyway* (=8.8) and *well* (=3.8) thus exemplify hybrid items that operate across contexts with varying pragmatic weight. Conversely, *indeed* demonstrates a clear academic preference, aligning with its established role in signaling epistemic reinforcement and rhetorical stance (Chung et al., 2023; Barbara et al., 2024). The high incidence of *indeed* in academic writing does highlight the preference of formal writing to position-oriented metadiscourse markers over interactional devices. This interpersonal/epistemic polarity of DMs supports the multidimensional register model of Biber, where spoken registers are involved and written registers are informationally dense.

The qualitative concordance analysis of *well* and *actually* shows the necessary subtleties in DM functioning. Although *well* demonstrated comparatively high proportions of discourse markers in spoken data (70%), in academic writing, it was only used in reported speech or parenthetical commentary. This proves that *well* is a relevant spoken organizational tool, which signifies the beginning of responses, fixing, or cushioning (Crible & Blackwell, 2020; Shustova & Tsarenko, 2019). The lower proportion of DM-like *well* in academic texts likely reflects genre-specific constraints that minimize interactional and evaluative markers in formal writing. By contrast, *actually* demonstrated low DM proportions in both registers (10% in spoken, 20% in academic). Its predominant truth-conditional function (e.g., "actually implemented," "actually work") affirms its lexical anchoring and supports prior claims that *actually* oscillates between propositional and metadiscursive uses (Traugott, 2021; Crible & Pascual, 2020). Interestingly, the occasional DM-like instances of *actually* in academic writing which are found often in quoted or contrastive environments suggest that when DMs do surface in academic texts, they do so under rhetorically marked conditions (e.g., argumentation, stance framing). Therefore, register variation in DM use involves not only frequency but functional specialization. Spoken discourse is more inclined towards intersubjective and sequential DMs, whereas written academic discourse is more inclined towards epistemic and textual markers, which is a trend that is observed in both cross-linguistic (Schnell & Schiborr, 2022) and pedagogical corpus (Le Foll, 2021, 2024).

The patterning of the observed registers is related to cross-linguistic research indicating that DM frequency is associated with orality and interactivity (Cuenca, 2022; Chapwanya & Nel, 2024; Apostol et al., 2025). Indicatively, Chapwanya and Nel (2024) found that *so* and *well* differ in frequency and pragmatic role in the Zimbabwean and British English, which highlights the role of sociocultural factors in DM implementation. In a similar manner, Apostol et al. (2025) have shown cross-linguistic differences in multiword markers (*you know*, *in fact*) based on genre and mode of communication. The existing results also overlap with the studies on polyfunctionality and co-occurrence patterns in spoken corpora (Crible & Cuenca, 2019), which proves that markers are frequently used in groups that represent common pragmatic functions instead of single cases. An example would be *you know* and *I mean*, which often occur together in speech as complementary position and clarification tools, an association which, perhaps, also explains their very high frequencies of speech.

In terms of cognitive-pragmatic perspective, DMs are observed to act as processing aids when spontaneous speech is produced. Their co-occurrence with disfluency characteristics (e.g., pauses, repairs) is an indication of the cognitive load of planning and monitoring speech (Crible, 2017). This practical position is the reason why they are relatively rare in written registers where the processing limitations are reduced to a minimum. The DM underrepresentation in scholarly prose also coincides with the rhetorical economy of formal writing, in which coherence is ensured by the use of explicit clause conjunctions (e.g., *therefore*, *however*) instead of interactive ones (Zhou & Xia, 2024; Barbara et al., 2024). These oppositions confirm that register does not only restrict the choice of language, but also organizes the very process of discourse.

The findings have a number of implications that can be made to inform theory and practice. Practically, the results support the need to incorporate register awareness and DM teaching in ELT programs. L2 learners tend to overgeneralize or underuse DMs because of the lack of exposure to real-life spoken input, as it has been demonstrated in numerous studies (Asik, 2015; Vickov & Jakupcevic, 2017). High-

frequency conversational DMs such as *you know* and *I mean* are usually underrepresented in pedagogical materials, with the focus on connective markers (*however, therefore*) that are characteristic of formal writing (Le Foll, 2021; Le Foll, 2024). Educators can develop pragmatic competence and assist students to adjust their discourse adequately in different contexts by providing them with corpus-based evidence of register-specific DM use. An example is that *you know* and *I mean* can be taught as signs of interactive conformity and clarification in spoken English, but *indeed* and *in fact* can be placed as formal signs of academic argumentation. This type of differentiation facilitates the understanding of situational suitability, which is a major constituent of communicative competence. Moreover, the functional validation process that we used in our research shows a pedagogically replicable model of classroom corpus tasks. Students were able to exercise concordance lines to differentiate between discourse and lexical applications of polyfunctional items, and this is an activity that has been demonstrated to enhance sensitivity to pragmatic subtlety and genre norms (Abidah et al., 2024; Cushing, 2017).

The current research also adds to the theoretical frameworks of register variation, pragmatic multifunctionality, and discourse organization. The findings quantitatively support the multidimensional model of Biber et al. (2002) by demonstrating that frequency differences between DMs are parallel to those between the involvement dimension and information density dimension. The data on DM proportions, in their functional aspect, indicate that the register differences cannot be represented by counting frequencies only; the functional coding is needed to distinguish between true discourse and lexical or grammaticalized versions (Crible & Pascual, 2020). The combination of normalized frequency analysis and a qualitative validation approach explains how corpus-based pragmatics can narrow down theoretical explanations of register. The data on the overlapping but distinct DM functions in registers confirms the opinion of Cuenca (2022) that DMs are interpersonal metaphors that are pragmatically adjusted to the situation, but not categories. Lastly, the paper is also a contribution to the field of integrating corpus linguistics and applied pragmatics, and how big data of empirical evidence can be interpreted pedagogically without losing theoretical rigor. COCA has a balanced design, which guarantees that the results represent general patterns of use and not genre-specific biases.

Although the current research has the advantage of scale and representativeness of COCA, some limitations are to be considered. To begin with, the KWIC validation sample size (20 lines per marker) provides suggestive rather than comprehensive information on functional proportions. A bigger, stratified concordance sample would permit more powerful statistical modeling of DM function. Second, the research is restricted to the American English language, which does not allow generalizing to other varieties of English; therefore, future studies might involve cross-corpora comparisons with the Global Web-Based English Corpus or the British National Corpus. In addition, the current analysis was limited to surface implementations of orthographic DMs. Multimodal and prosodic aspects, particularly those of interest to spoken markers such as *well* and *you know*, may also shed more light on how DMs mark discourse structure. Lastly, longitudinal and learner corpora studies might follow the development of DM use at different levels of proficiency, and this would add to the existing EFL-oriented studies (Neary-Sundquist, 2014; Asik, 2015).

Conclusion

This paper affirms that the use of discourse markers in the English language is highly influenced by register. Spoken discourse has a strong dependence on interactive markers (*you know, I mean, well*) that can control interpersonal alignment and real-time processing, whereas academic writing has a strong dependence on epistemic and rhetorical markers (*indeed, in fact*) which can control argumentation. Polyfunctional markers such as *well* and *actually* emphasize the necessity of functional validation of the raw frequency counts since their application in discourse is extremely context-related. The results are consistent with the current theories of pragmatic variation and elaborate on them by measuring the level of register differentiation in a large modern corpus. Pedagogically, they focus on the need to teach DMs as register-bound resources and not as universal connectors. Further studies in multimodal and cross-varietal data in the future will help us better understand the functioning of DMs in the varied communicative environments of English.

REFERENCES

- Abidah, K. H., Ratih, E., Zen, E. L., & Lumbanbatu, I. M. F. (2024). Understanding how ideas are linked in compare and contrast essays: A discourse marker (Dm) study. *REiLA : Journal of Research and Innovation in Language*, 6(3), 280–290. <https://doi.org/10.31849/reila.v6i3.15636>
- Apostol, E., Trucă, C., Damova, M., Silvano, P., Oleškevičienė, G. V., Liebeskind, C., Trajanov, D., Baczkowska, A., Montecchiari, E. A., & Chiarcos, C. (2025). Multiword discourse markers across languages: A linguistic and computational perspective. *International Journal of Applied Linguistics*, 35(4), 2078–2090. <https://doi.org/10.1111/ijal.12755>
- Aşık, A. (2015). Discourse markers in EFL setting: Perceptions of Turkish EFL teachers. *Educational Research and Reviews*, 10(7), 941–950. <https://doi.org/10.5897/ERR2015.2154>
- Barbara, S. W. Y., Afzaal, M., & Aldayel, H. S. (2024). A corpus-based comparison of linguistic markers of stance and genre in the academic writing of novice and advanced engineering learners. *Humanities and Social Sciences Communications*, 11(1), 284. <https://doi.org/10.1057/s41599-024-02757-4>
- Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press.
- Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, 36(1), 9. <https://doi.org/10.2307/3588359>
- Chapwanya, F. C., & Nel, J. H. (2024). Discourse markers *so* and *well* in Zimbabwean English: A corpus-based comparative analysis. *World Englishes*, 43(1), 23–46. <https://doi.org/10.1111/weng.12626>
- Chung, E., Crosthwaite, P. R., & Lee, C. (2023). The use of metadiscourse by secondary-level Chinese learners of English in examination scripts: Insights from a corpus-based study. *International Review of Applied Linguistics in Language Teaching*, 62(2), 977–1008. <https://doi.org/10.1515/iral-2022-0155>
- Crible, L. (2017). Discourse markers and (Dis)fluency in English and French: Variation and combination in the DisFrEn corpus. *International Journal of Corpus Linguistics*, 22(2), 242–269. <https://doi.org/10.1075/ijcl.22.2.04cri>
- Crible, L., & Blackwell, S. E. (2020). Introduction: Discourse-pragmatic markers in speech and sign. *Journal of Pragmatics*, 156, 24–27. <https://doi.org/10.1016/j.pragma.2019.09.002>
- Cuenca, M. J., & Crible, L. (2019). Co-occurrence of discourse markers in English: From juxtaposition to composition. *Journal of Pragmatics*, 140, 171–184. <https://doi.org/10.1016/j.pragma.2018.12.001>
- Cuenca, M.-J. (2022). Language norm and usage change in Catalan discourse markers: The case of contrastive connectives. *Languages*, 7(1), 66. <https://doi.org/10.3390/languages7010066>
- Cushing, S. T. (2017). Corpus linguistics in language testing research. *Language Testing*, 34(4), 441–449. <https://doi.org/10.1177/0265532217713044>
- Engel, A., Grafmiller, J., Rosseel, L., & Szmrecsanyi, B. (2022). Assessing the complexity of lectal competence: The register-specificity of the dative alternation after *give*. *Cognitive Linguistics*, 33(4), 727–766. <https://doi.org/10.1515/cog-2021-0107>
- Furko, P. (2017). Manipulative uses of pragmatic markers in political discourse. *Palgrave Communications*, 3(1), 17054. <https://doi.org/10.1057/palcomms.2017.54>
- Jucker, A. H. (2021). Features of orality in the language of fiction: A corpus-based investigation. *Language and Literature: International Journal of Stylistics*, 30(4), 341–360. <https://doi.org/10.1177/09639470211047751>
- Le Foll, E. (2021). Register variation in school EFL textbooks. *Register Studies*, 3(2), 207–246. <https://doi.org/10.1075/rs.20009.lef>

- Le Foll, E. (2024). Schulenglisch: A multi-dimensional model of the variety of English taught in German secondary schools. *Arbeiten aus Anglistik und Amerikanistik*, 49(1). <https://doi.org/10.24053/AAA-2024-0001>
- Neary-Sundquist, C. (2014). The use of pragmatic markers across proficiency levels in second language speech. *Studies in Second Language Learning and Teaching*, 4(4), 637. <https://doi.org/10.14746/ssllt.2014.4.4.4>
- Schnell, S., & Schiborr, N. N. (2022). Crosslinguistic corpus studies in linguistic typology. *Annual Review of Linguistics*, 8(1), 171–191. <https://doi.org/10.1146/annurev-linguistics-031120-104629>
- Shirtz, S. (2024). Discourse markers as the locus of signaling the main-event line in Alsea narratives. *Linguistics*, 62(1), 229–260. <https://doi.org/10.1515/ling-2021-0038>
- Shustova, S. V., & Tsarenko, N. M. (2019). Pragmatic potential of discourse markers. *Theoretical and Applied Linguistics*, 5(4), 163–170. https://doi.org/10.22250/2410-7190_2019_5_4_163_170
- Szmrecsanyi, B. (2019). Register in variationist linguistics. *Register Studies*, 1(1), 76–99. <https://doi.org/10.1075/rs.18006.szm>
- Tarantini, A. T., & Benatti, R. (2021). Translating discourse markers in theater: David Mence's *Convincing Ground* in Italian translation. *Translation and Interpreting Studies*, 16(1), 80–100. <https://doi.org/10.1075/tis.18030.tar>
- Traugott, E. C. (2021). A constructional perspective on the rise of metatextual discourse markers. *Cader-nos de Linguística*, 2(1), 01–25. <https://doi.org/10.25189/2675-4916.2021.v2.n1.id269>
- Vickov, G., & Jakupčević, E. (2017). Discourse markers in non-native EFL teacher talk. *Studies in Second Language Learning and Teaching*, 7(4), 649–671. <https://doi.org/10.14746/ssllt.2017.7.4.5>
- Zhou, J., & Xia, Y. (2024). Register-based distribution of expressions of modality in COCA. *Humanities and Social Sciences Communications*, 11(1), 650. <https://doi.org/10.1057/s41599-024-03134-x>