

CONTRASTIVE CORPUS STUDIES IN LANGUAGE TEACHING: A BIBLIOMETRIC AND SYSTEMATIC REVIEW OF LANGUAGE PAIRS

¹*Yulia Yulian, ¹Yeti Mulyati, ¹Ida Widia

¹Faculty of Language and Literature Education, Universitas Pendidikan Indonesia, Jl. Dr. Setiabudi No.229, Isola, Kec. Sukasari, Kota Bandung, West Java, Indonesia

*Corresponding Author Email: yuliyulian2024@upi.edu

Article Info	Abstract
Article History Received: April 2025 Revised: June 2025 Accepted: September 2025 Published: October 2025	<i>Corpus-based contrastive studies have become increasingly vital in language education, offering empirical insights into cross-linguistic structures, discourse patterns, and semantic strategies. However, a comprehensive synthesis mapping recent developments and identifying existing research gaps has been lacking. This study addresses this need by conducting a systematic literature review and bibliometric analysis of 73 peer-reviewed articles published between 2014 and 2024. Using PRISMA 2020 guidelines and tools such as VOSviewer, the study maps dominant linguistic focuses, frequently analyzed language pairs, and collaboration networks. The findings reveal a strong emphasis on discourse and semantics, with English commonly paired with Spanish, Arabic, and Chinese. Pragmatic dimensions and underrepresented language pairs, particularly from Southeast Asia and Africa, remain largely overlooked. The study also highlights geographic disparities in global contributions, suggesting the need for more inclusive and collaborative research frameworks. The results offer practical implications for pedagogy, policy, and research. For pedagogy, integrating corpus-informed, contrastive materials into language instruction can enhance learners' awareness of discourse conventions and intercultural variation. In terms of policy, stakeholders are encouraged to invest in open-access multilingual corpora and support cross-border collaboration, particularly involving underrepresented linguistic communities. For future research, this study underscores the need to explore pragmatic features, diversify language pairings, and build interdisciplinary networks to promote a more equitable and context-sensitive development of corpus-based contrastive inquiry in language education.</i>
Keywords Bibliometric analysis; Contrastive analysis; Corpus; Language pair; Systematic review;	
How to cite: Yulian, Y., Mulyati, Y., & Widia, I. (2025). Contrastive Corpus Studies in Language Teaching: A Bibliometric and Systematic Review of Language Pairs. <i>JOLLT Journal of Languages and Language Teaching</i> , 13(4), 1994-2007. Doi: https://doi.org/10.33394/joltt.v13i4.15236	

INTRODUCTION

Contrastive studies play a pivotal role in foreign language teaching by highlighting both the similarities and differences between languages and cultures, thereby enhancing learners' understanding of their native and target languages (Moinzadeh, Dezhara, & Rezaei, 2012). These insights are especially valuable for educators, as they enable the anticipation of common learner difficulties that often stem from first-language interference (Brown, 2008; (Brown, 2008; Jafari, 2014). Contrastive analysis also fosters learners' linguistic awareness by helping them recognize how cross-linguistic influence contributes to language errors (James, 2005; Mair, 2005). In addition to diagnosing error patterns, contrastive studies support the development of pedagogical grammars, the selection of effective teaching materials, and the design of adaptive instructional strategies tailored to learners' needs (Gass & Selinker, 2008; Moinzadeh, Dezhara, & Rezaei, 2012). Moreover, when integrated with corpus linguistics and semantic analysis, contrastive analysis becomes more empirically

robust, providing deeper insights into authentic language use (Juodinytė-Kuznetsova, 2015; Wu & Baccanello, 2019). Such findings not only inform general language pedagogy but also serve as a foundation for developing contextually relevant materials (Pablos-Ortega, 2015).

With the rise of digital corpora and analytical tools such as SketchEngine and AntConc (Anthony, 2013), corpus linguistics has become an indispensable method in contrastive studies. It enables researchers to access, process, and analyze large linguistic datasets (Qin, 2015)(Xie, 2015), and has been widely applied in language testing, translation studies, and L2 instruction (Altameemi, 2024). Beyond structural comparisons, corpus-based approaches allow researchers to explore discourse, pragmatics, and sociocultural dimensions (Aijmer, 2020; Andersen, 2023). They also support data-driven comparisons using translation equivalence as a *tertium comparationis* (García, 2021; Lyu & Wang, 2015), thus broadening the scope of contrastive analysis across more diverse languages and topics (Hasselgård, 2020), and supporting the development of translation studies as a descriptive discipline (Ibrahim, 2015).

Despite growing scholarly interest in corpus-based contrastive studies, comprehensive syntheses that systematically map the empirical development of the field remain scarce. The most substantial contribution thus far, Hasselgård's (2020) review, offers valuable historical perspectives, particularly on the emergence of parallel corpora in Scandinavia, and addresses key methodological and theoretical issues. However, its scope is confined to studies published up to 2018 and lacks a systematic account of the linguistic levels, contrastive aspects, and language pair distributions examined in more recent research. Two other recent reviews deserve mention. Altameemi's (2024) state-of-the-art review provides a broad overview of corpus linguistics development, from foundational figures to recent innovations such as ChatGPT, offering relevant insights into the field's methodological evolution and technological trajectories. However, while useful in contextualizing corpus linguistics as a whole, the review does not specifically address contrastive studies or examine linguistic pairing trends. Similarly, Lin & Lei's (2020) bibliometric analysis of multilingualism in applied linguistics and education is highly relevant from a contrastive perspective. Nevertheless, their study does not focus on corpus-based contrastive methodologies or provide detailed mapping of language pair distributions and linguistic levels.

In light of rapid advancements in corpus technologies and the increasing diversification of analytical approaches, the landscape of corpus-based contrastive studies has likely undergone considerable transformation over the past decade. Yet, no updated review has been conducted to document these changes or provide an evidence-based understanding of the field's current state. To address this gap, the present study adopts a data-driven approach by integrating a Systematic Literature Review (SLR) with bibliometric analysis to examine corpus-based contrastive studies published between 2014 and 2024. To guide this investigation, the study addresses the following research questions: what linguistic levels and contrastive aspects are most frequently examined in corpus-based contrastive studies in language teaching?; which language pairs have been most and least represented in these studies over the past decade?; and what are the bibliometric characteristics of the field, including publication trends, leading contributors, collaboration networks, and commonly used keywords?

RESEARCH METHOD

Research Design

This study employed a mixed-methods approach by integrating a Systematic Literature Review (SLR) with bibliometric analysis to examine the development of corpus-based contrastive studies in language teaching from 2014 to 2024. The qualitative component involved a thematic synthesis of linguistic levels (e.g., syntax, pragmatics, discourse), contrastive aspects, and language pair combinations. The quantitative component applied

bibliometric analysis to examine annual publication trends, influential contributors, citation metrics, collaboration networks, and keyword patterns. The entire review process was guided by the PRISMA 2020 protocol (Page et al., 2021), ensuring transparency, replicability, and methodological rigor. This integrated approach allows for a comprehensive mapping of empirical and conceptual developments in the field while supporting evidence-based directions for future research.

Research Subjects

The research subjects consisted of peer-reviewed journal articles focused on corpus-based contrastive studies in the context of language teaching. A purposive sampling method was employed, and the four-stage PRISMA process was used to ensure transparency and replicability:

Step 1: Identification

Relevant articles were retrieved from Scopus using Boolean operators and exact phrases. The search string used was: ("contrastive analysis" OR "corpus-based contrastive study" OR "comparative corpus analysis"). This query yielded 3,275 documents.

Step 2: Screening

Titles, abstracts, and full texts were screened manually. Articles were evaluated against predefined inclusion and exclusion criteria (see Table 1), and duplicates or irrelevant entries were removed.

Table 1
Inclusion and Exclusion Criteria

Category	Inclusion Criteria	Exclusion Criteria
Publication Type	Peer-reviewed journal articles	Grey literature, review, theses, conference proceedings, and non-peer-reviewed sources
Research Focus	Studies using corpus-based contrastive approaches comparing two or more languages	Studies focusing solely on corpus linguistics, translation, error analysis, or interlanguage
Linguistic Aspect	Studies that explicitly compare linguistic levels or features	Studies focusing only on monolingual acquisition of linguistic features
Language	Articles published in English	Articles published in languages other than English
Publication Year	Published between 2014–2024	Published outside of the 2014–2024 range
Access	Open-access or institutionally accessible articles	Closed-access articles without full-text availability

Step 3: Data Extraction

Selected articles ($n = 73$) were compiled and organized in Google Sheets, which facilitated the extraction of metadata (e.g., author, year, journal, linguistic focus, method, language pair).

Step 4: Synthesis and Interpretation

Articles were classified thematically by linguistic domain and contrastive features. Bibliometric analysis was used to map publication patterns, co-authorships, and keyword trends.

Instruments

The instruments used were: (1) the Scopus database (Elsevier) for retrieving peer-reviewed publications relevant to corpus-based contrastive studies; (2) Google Sheets for organizing and managing extracted metadata (author, year, title, linguistic focus, language pair, etc.); and (3) VOSviewer software (version 1.6.20, Van Eck & Waltman, Leiden

University) for generating bibliometric visualizations including co-authorship networks, keyword co-occurrence maps, and citation patterns. The first author conducted all search, screening, and coding procedures manually. While intercoder reliability testing was not applied, all decisions were documented and verified through multiple rounds of self-checking to enhance consistency. The synthesis phase incorporated bibliometric mapping based on (Donthu, Kumar, Mukherjee, Pandey, & Lim, 2021), focusing on performance metrics, citation analysis, and keyword network visualization.

Data Analysis

Both qualitative and quantitative data analysis methods were employed. To address the research questions, a thematic synthesis was conducted following qualitative data coding principles proposed by (Skjott Linneberg & Korsgaard, 2019). A primarily deductive coding approach was used, with pre-defined codes based on theoretical categories derived from the research questions and literature (e.g., discourse, syntax, metaphor, modality). During the coding process, if relevant concepts not initially included emerged, inductive codes were added to capture those insights. This hybrid approach allowed for both structure and openness to emerging themes. Coding was conducted manually in Google Sheets by the first author. Although intercoder reliability testing was not conducted, the lead researcher repeated the coding process in three cycles, using cross-checking and memoing to enhance reliability and maintain analytical consistency.

VOSviewer was used to generate visual representations of collaboration networks, citation relationships, and keyword trends. Frequency counts and citation metrics were normalized using default settings in VOSviewer, and the visualizations were interpreted in relation to the research questions and thematic findings. While the scope of this review was limited to Scopus-indexed, English-language articles, efforts were made to include studies from diverse geographical regions and linguistic contexts. Limitations related to database coverage and single-coder analysis are acknowledged and discussed in the conclusion.

RESEARCH FINDINGS AND DISCUSSION

Research Findings

Emerging Trends in Corpus-Based Contrastive Studies: A Decade of Growth (2014–2024)

Figure 1 illustrates the annual publication trend of corpus-based contrastive studies from 2014 to 2024, based on data retrieved from the Scopus database.

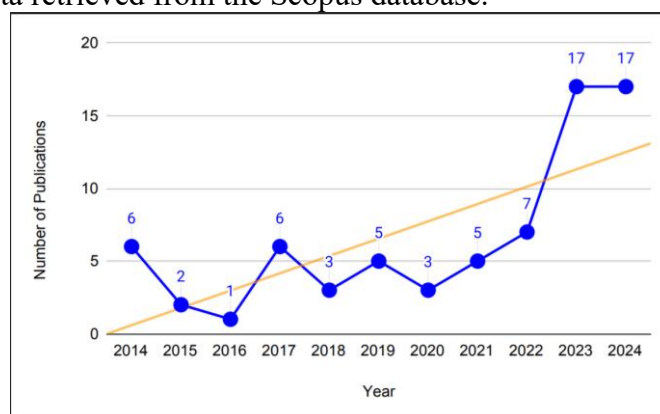


Figure 1. Annual Progress in Corpus-Based Contrastive Studies

The overall trajectory of corpus-based contrastive studies reflects both quantitative and qualitative growth over the past decade. As shown in Figure 1, the number of publications remained relatively stable between 2014 and 2020, with a slight decline during 2015–2016, followed by a gradual rise after 2017. A significant increase occurred after 2021, culminating in the highest publication output in 2023 and 2024, with 17 studies each year. This upward trend aligns with the evolving scholarly focus and expanding scope of the field. In the early

2010s, research primarily concentrated on grammatical and lexical structures, such as Jafari's (2014) analysis of English–Persian prepositions and Van Goethem & De Smet's (2014) study on how noun turns to adjective in French, English and Dutch. By 2015, attention had shifted toward syntactic constructions and grammaticalization (e.g., De Cesare & Garassino (2015); Leuschner & Van Den Nest (2015)). The period from 2016 to 2019 marked a growing interest in discourse and genre analysis, as seen in studies on discourse markers (Crible, Degand, & Gilquin, 2017) and cross-linguistic genre comparisons (Arroyo & Roberts, 2017).

The Mapping of Language Pairs in Corpus-Based Contrastive Research

To gain a clearer picture of current research priorities in corpus-based contrastive studies, it is essential to examine which language pairs have received the most scholarly attention over the past decade. Table 2 presents a sample of the most frequently studied language pairs, selected from the top-ranking entries in the full dataset. This snapshot highlights dominant combinations as well as the relative scarcity of certain pairs, offering insight into prevailing research trends and potential gaps in the field.

Table 2
Language Pair Frequencies in Corpus-Based Contrastive Studies (2014–2014)

Language Pair	Frequency	Author (Year)
English–Spanish	6	Izquierdo, M., & Pérez Blanco, M. (2023); Lastres-López, C., & Camacho-Salas, N. (2023); Pérez-Ruiz, L. (2024); Elorza, I. (2014); Arroyo, B. L., & Roberts, R. P. (2017); Bellés-Fortuño, B. (2017)
English–Arabic	5	Tarawneh, R. T., & Al-Momani, I. M. (2023); Elgobshaw, A. E. (2024); Al-Khaza'leh, B. A., & Alzubi, A. A. F. (2022); Al-Ajlouni, S. A. (2024); Joodi, A. M. H. (2023)
English–Chinese	5	Wang, F., Jia, X., & Lin, Z. (2023); Gardner, S., & Han, C. (2018); Cheng, C. (2014); Li, B. (2014); Li, Y. (2020)
English–French	4	Curry, N., & Chambers, A. (2017); van Wette, N. (2021); Crible, L., Degand, L., & Gilquin, G. (2017); Cappelle, B., Mostrov, V., & Tayalati, F. (2021)
English–Indonesian	3	Auni, L., & Manan, A. (2023); Dewiyanti, S., Yuliyawati, S. N., Saudin, & Meilinda, L. (2024); Al Fajri M.S.; Oktavianti I.N. (2024)
English–Vietnamese	2	Hong, N. B., & Pham, H. (2024); Vu, M. P. (2024)
English–German	2	Leuschner, T., & Van Den Nest, D. (2015); Podobnik, A. (2021)
English–Italian	2	De Cesare, A.-M., & Garassino, D. (2015); Molino, A. (2017)
Albanian–Serbian	1	Sejdiu-Rugova, L., & Krijezi, M. (2018)
American English & Malaysian English	1	Zarza, S., & Tan, H. (2016)

As can be seen in Table 2, corpus-based contrastive studies show a strong tendency to focus on English as the anchor language, particularly in combination with globally spoken or geopolitically significant languages. The most frequent language pair is English–Spanish, appearing in six studies, followed by English–Arabic and English–Chinese (five studies each), and English–French (four studies). These combinations suggest a prioritization of language pairs with high communicative reach, pedagogical value, or established academic tradition. In contrast, studies on English–Indonesian, English–Vietnamese, English–German, and English–Italian are less frequent, each appearing in only two to three studies, despite the relevance of these pairs in regional multilingual settings and emerging academic environments. A few language pairs, such as Albanian–Serbian and American English–

Malaysian English, along with others lower in the ranking list, appear only once, reflecting more idiosyncratic or localized research interests.

Dominant Linguistic Aspects in Corpus-Based Contrastive Studies

A closer look at the linguistic dimensions explored in recent studies reveals notable patterns, as summarized in Table 3, which categorizes the focus areas based on their frequency.

Table 3
Frequency of Linguistic Levels and Aspects in Corpus-Based Contrastive Studies (2014–2024)

Linguistic Domain	Linguistic Aspect	Frequency
Discourse	Lexicogrammatical strategies; Spoken discourse patterns; Editorial structure and strategy; Discourse markers and pauses (fluency); Genre and register; Evaluative expressions; Contrastive connectors; Reformulation markers; Theme dynamics in forums; Expressions of aggressiveness in written discourse; Metadiscourse strategies (e.g., engagement markers); Deverbal discourse-pragmatic markers (wait, espera); Judicial decision texts; Lexical repetition; Lexical bundles in test passages; Rhetorical organization of texts; Disinformation discourse; Reformulation strategies across modalities; Persuasive rhetorical strategies; Discursive strategies in climate-related texts; Stance markers in academic writing	22
Semantics	Attitudinal meanings (Appraisal theory); Idiomatic expressions; Idiomatic expressions; Minimizer/maximizer idioms; Color term meanings; Semantic/constructional encoding of traits; Somatic idioms (phraseological units containing body parts); Wish formulae (idiomatic expressions for wishing good luck); Conceptual metaphor of “love”; Lexical/semantic translation of ‘marsh’ terms; Figurative comparisons using animals; Metaphorical collocations; Spatial expressions and orientation; Semantic opposites in proverbs; Idiomatic similes; Conceptual opposition of wealth/poverty; Semantic and idiomatic use of face; Collocational mapping; Emotional concepts comparison; Deictic motion and anchoring; Conceptual framing and reframing	21
Syntax	Prepositions; Passive construction; Cleft constructions and exhaustiveness; V1-conditionals; Interrogative sentence structures; Reporting clauses; Double-object construction; Passivity through copular constructions; Ergativity (agent-patient marking); Closed interrogative sentence structures; Hyphenated premodifiers	11
Morphology	Debonding (noun-to-adjective shift); Feminine forms (gendered nouns); Dual expressions; Word structure (morphological comparison); Past future perfect tense; Terminological definitions of parts of speech; Grammatical aspect (perfective/imperfective); Gerund usage; Morphological root comparison; Suffix (-ish) in informal writing	10
Morphology & Syntax	Copula verbs ("to be"); Adjective formation and structure; Modal verbs; Participles and gerunds; Past imperfect tense; Influence of Greek grammar on early French grammar description	6
Pragmatics	Greeting expressions; Invitations; Illocutionary indicators in questions	3

Based on the Table 3, it can be seen that over the past decade, corpus-based contrastive studies have placed significant emphasis on discourse and semantic levels of linguistic analysis. As indicated in the data, discourse-level features are the most frequently examined (22 instances), encompassing a wide range of concerns such as lexicogrammatical strategies (Elorza, 2014), discourse markers (Crible et al., 2017), metadiscourse (Izquierdo & Pérez Blanco, 2023), genre (Arroyo & Roberts, 2017), disinformation discourse (Ruzaité,

2024), and rhetorical organization (Hong & Pham, 2024)—reflecting sustained scholarly interest in how language is structured and used in context. Closely following is the semantic level (21 instances), with studies exploring idiomatic expressions (Blagoeva, Jaskot, & Sosnowski, 2020), conceptual metaphors (Al-Khaza'leh & Alzubi, 2022), appraisal meanings (Cheng, 2014), and various figurative or culturally embedded lexical patterns.

Syntax and morphology received moderate attention (11 and 10 instances respectively), focusing on structural contrasts such as clefting (De Cesare & Garassino, 2015), interrogatives (e.g. Curry & Chambers (2017)), word formation (e.g. Sareh (2023)), and morphological encoding across languages (e.g. Elgobshawy (2024)). Combined morphology-syntax features (6 instances), including modal verbs (Mohd Jalis & Abd Rahim, 2014) and participial constructions (Šimůnková, 2018), also appear, though less frequently. Notably, pragmatic aspects remain underexplored (3 instances), limited to studies on greetings, invitations, and illocutionary indicators (Curry, 2023; Elhami, 2020; Kotorova, 2021).

Author Collaboration (Co-Authorship) Map in Corpus-Based Contrastive Studies

Based on the data analysis in VosViewer, there are 114 authors across 73 publication documents related to corpus-based contrastive studies, as shown in Figure 2.

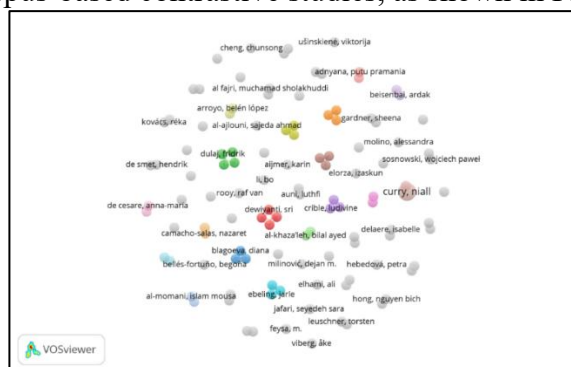


Figure 2. Network Visualization of Co-Authorship in Corpus-Based Contrastive Studies (2014–2024)

Figure 2 visualizes co-authorship patterns in corpus-based contrastive studies. Each node represents an author, with connecting lines indicating collaborations. Distinct color-coded clusters suggest thematic, institutional, or regional ties. Niall Curry appears central, actively engaging in multiple co-authorships, while Sri Dewiyanti, Fridrik Dulaj, Ludivine Crible, and Diana Blagoeva serve as key connectors. Peripheral nodes indicate limited collaboration or emerging involvement.

Global Distribution in Corpus-Based Contrastive Studies

Figure 3 below visualizes the global distribution of co-authorship in corpus-based contrastive studies from 2014 to 2024.

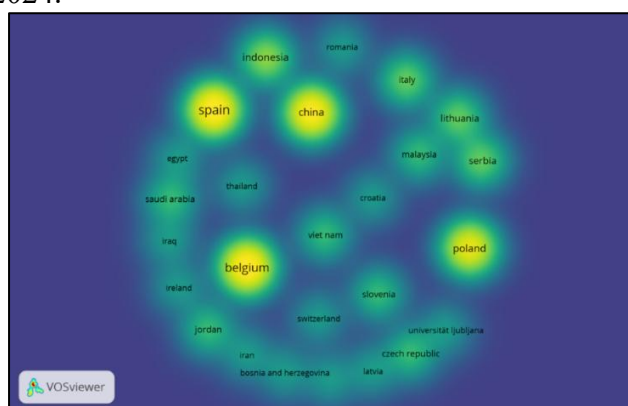


Figure 3 Global Distribution in Corpus-Based Contrastive Studies (2014–2024)

Based on Figure 3, Spain, China, Belgium, and Poland emerge as the most prolific contributors, suggesting strong institutional support and access to multilingual corpora. Spain stands out as the most central hub, followed by China and Belgium, which also show significant co-authorship intensity. Poland, though slightly less central, displays a focused and consistent output. Medium-density clusters are visible in countries like Indonesia, Malaysia, Lithuania, Serbia, and Italy, reflecting growing regional involvement. Indonesia's presence, in particular, signals Southeast Asia's rising engagement in global contrastive linguistics research. Meanwhile, lighter activity appears in Vietnam, Switzerland, Slovenia, Iran, and Jordan—indicating emerging participation or limited access to collaborative networks.

Keyword Map in Corpus-Based Contrastive Studies

The following figure, Figure 4, visualizes the keyword co-occurrence network in corpus-based contrastive studies, highlighting the thematic connections and research trends within the field over the past decade.

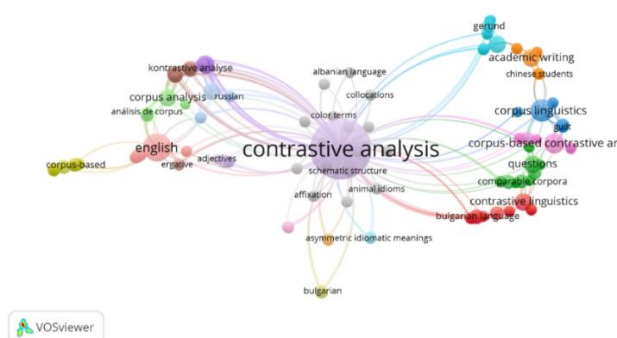


Figure 4. Network Visualization of Keyword Co-Occurrence in Corpus-Based Contrastive Studies (2014–2024)

Figure 4 reveals a thematically cohesive and methodologically diverse research landscape, with “contrastive analysis” occupying the central position. Surrounding clusters represent key thematic areas, including methodological focuses such as “contrastive linguistics” and “comparable corpora”, and “corpus linguistics” applied educational contexts involving “Chinese students”, and “academic writing”, as well as fine-grained linguistic analyses of features like “gerund”, “collocations”, “affixation”, and “idiomatic meanings”. The inclusion of multilingual terms such as “kontrastive analyse” and “análisis de corpus” points to broad international contributions. Strong co-occurrence links reflect an integrated field that connects applied linguistics, language pedagogy, and typological comparison, while offering room for further exploration of less-studied language pairs and culturally grounded contrastive frameworks.

Highly Cited Works in Corpus-Based Contrastive Studies

The Table 1 below presents the most cited works in corpus-based contrastive studies over the past decade.

Table 4
Most Cited Corpus-Based Contrastive Studies (2014–2024)

Author (Year)	Title	Citation
Crible (2017)	The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis)fluency	38
Van Goethem (2014)	How nouns turn into adjectives: The emergence of new adjectives in French, English and Dutch through debonding processes	19
Gardner (2018)	Transitions of contrast in Chinese and English university student writing	17

Author (Year)	Title	Citation
De Cesare (2015)	On the status of exhaustiveness in cleft sentences: An empirical and cross-linguistic study of English also-/only-clefts and Italian anche-/solo-clefts	12
Sosnowski (2019)	A contrastive analysis of feminitives in Bulgarian, Polish and Russian	10
Curry & Chambers (2017)	Questions in English and French Research Articles in Linguistics: A Corpus-Based Contrastive Analysis	9
Leuschner (2015)	Asynchronous grammaticalization: V1-conditionals in present-day English and German	9
Zarza (2016)	Patterns of schematic structure and strategic features in newspaper editorials: A comparative study of American and Malaysian editorials	9
Elhami (2020)	A socio-pragmatic perspective of Spanish and Persian greeting	8
Curry (2023)	Question illocutionary force indicating devices in academic writing A corpus-pragmatic and contrastive approach to identifying and analyzing direct and indirect questions in English, French, and Spanish	7
Arroyo (2017)	Genre and register in comparable corpora: An English/Spanish contrastive analysis	6

Table 4 highlights patterns in thematic focus, methodology, and scholarly relevance. Crible et al. (2017), which leads the list, explores the clustering of discourse markers and filled pauses in French and English, indicating a strong academic interest in pragmatic fluency and spoken language features. Similarly, Curry & Chambers (2017) and De Cesare & Garassino (2015) focus on interrogatives and cleft constructions, respectively, reflecting a preference for contrastive investigations of syntactic and information-structuring devices. Morphosyntactic topics such as debonding (Van Goethem & De Smet, 2014) and grammaticalization processes (Leuschner & Van Den Nest, 2015) also appear to attract sustained scholarly attention, suggesting their theoretical importance across languages. Meanwhile, studies incorporating sociocultural or register-sensitive elements, such as Sosnowski & Satoła-Staskowiak (2019) on feminitives and Zarza & Tan (2016) on editorial discourse—demonstrate the interdisciplinary appeal of contrastive corpus research.

In contrast, several studies within the same corpus-based contrastive category have received zero citations. Despite their linguistic diversity and contrastive analytical approaches, they share several characteristics that may contribute to their current lack of citations. Most of these studies are highly specialized, focusing on narrow grammatical, lexical, or pragmatic phenomena across less globally dominant languages, such as the use of the past imperfect tense in Albanian and Serbian (Sejdiu-Rugova & Krijezi, 2018), wealth/poverty oppositions in English and Kazakh (Kerimbayeva & Beisenbai, 2024), or deixis in Estonian, Finnish, and Czech (Hebedová, 2024), which may limit their broader academic appeal or interdisciplinary relevance. Additionally, many of them are rooted in traditional contrastive analysis without integrating current theoretical frameworks or applied linguistic perspectives.

Discussion

The findings of this study reaffirm and significantly extend current understandings of corpus-based contrastive studies in language teaching by situating them within a broader shift in applied linguistics toward data-driven, multilingual, and contextually grounded research. The ten-year bibliometric trend from 2014 to 2024 reveals a substantial increase in scholarly output, particularly after 2021, which suggests a marked reorientation in academic priorities. This trajectory aligns with the evolving landscape of applied linguistics that increasingly

prioritizes empirical rigor and contextual sensitivity in language analysis. Recent scholarship characterizes this period as a “golden age” for the field, marked by a growing emphasis on multilingual comparative inquiry, sociocultural dimensions of language use, and identity-related issues (McKinley, 2019). These developments collectively indicate that corpus-based contrastive studies are not only benefiting from this broader intellectual momentum but are also contributing meaningfully to the field’s theoretical and methodological innovation. These developments also carry important implications for future research, pedagogy, and language policy, which are further elaborated below.

The dominance of English-based pairings, especially English–Spanish, English–Arabic, and English–Chinese, reflects global communicative priorities and unequal access to linguistic resources and infrastructure, a pattern also noted in previous studies (Hasseltgård, 2020; Lyu & Wang, 2015). This linguistic centrality of English mirrors its status as a *lingua franca* in both academic and educational contexts (Jenkins, 2013), but also reveals geographic and typological imbalances in research coverage. Southeast Asian, African, and minority language pairs remain critically underrepresented. This lack of inclusivity hinders the global pedagogical applicability of contrastive insights and raises questions about equity in linguistic knowledge production. This uneven representation contrasts with the findings of Lei & Liu (2019), who observed an increasing visibility of contributions from non-traditional publishing countries in applied linguistics, reflecting a more globally distributed knowledge production and signaling a paradigm shift toward greater inclusivity and diversification in the field.

Regarding linguistic domains, the predominant focus on discourse and semantic levels observed in this study aligns with prior findings by McKinley (2019) and Lei and Liu (2019), who identified such trends as key characteristics of what has been termed the “golden age” of applied linguistics. This era is also marked by a noticeable shift away from traditional concerns such as phonology and syntax, and toward more context-sensitive approaches that prioritize meaning-making and functional language use across genres and cultures. The increasing emphasis on language in use underscores the importance of discourse-level investigations, which facilitate the analysis of genre-specific structures, metadiscourse elements, and rhetorical strategies that shape interpretation and learner production. Similarly, semantic-oriented studies offer valuable insights into cross-linguistic conceptual transfer and cultural framing (Charteris-Black, 2004; Kövecses, 2002). At the same time, the field has witnessed a concurrent expansion into the cognitive and psycholinguistic dimensions of multilingualism (Lin & Lei, 2020), reflecting a broader theoretical diversification. Nevertheless, pragmatic aspects remain significantly underexplored.

The co-authorship and global distribution maps reveal a core-periphery dynamic in scholarly collaboration. Consistent with findings by Hyland (2016) and Lei & Liu (2019), this study shows that scholars from Europe, China, and select Anglophone contexts dominate the field, while researchers from the Global South—especially Sub-Saharan Africa and parts of Southeast Asia—remain largely peripheral. This pattern not only reflects disparities in institutional support, access to corpora, and publication networks but also suggests missed opportunities for knowledge exchange and capacity building. International collaboration, corpus-sharing, and regional corpus development are needed to balance this asymmetry and enrich typological diversity. This has direct implications for research strategy and infrastructure: encouraging more inclusive co-authorship networks and open-access corpus initiatives could foster innovation and reduce epistemic inequality in applied linguistics.

The keyword co-occurrence map further supports the claim that the field is both thematically cohesive and methodologically expansive. “Contrastive analysis” remains the dominant node, indicating its continued centrality in the field. Surrounding clusters such as “comparable corpora”, “academic writing”, and “Chinese students” illustrate how corpus-based contrastive research is increasingly applied to pedagogical and learner-centered

contexts. These findings resonate with trends observed by Römer & Wulff (2010), who emphasize the role of learner corpora and cross-cultural academic discourse in shaping applied corpus research. Moreover, the presence of fine-grained linguistic features such as “idiomatic meanings” and “collocations” reflects a growing interest in meaning-based and usage-oriented analyses, which are essential for designing effective instructional materials and translation frameworks (Granger, Gilquin, & Meunier, 2015). For language educators and curriculum developers, this suggests the growing relevance of corpus-informed resources to support vocabulary acquisition, register awareness, and contrastive pragmatics in language classrooms.

CONCLUSION

This study offers a comprehensive, data-driven synthesis of corpus-based contrastive research in language teaching from 2014 to 2024, revealing significant growth in publication volume, thematic diversity, and scholarly attention—particularly in English-paired language studies such as English–Spanish, English–Arabic, and English–Chinese, with a strong emphasis on discourse and semantic analysis. Despite this progress, critical gaps persist: underrepresentation of Southeast Asian, Eastern European, and African languages, limited engagement with pragmatic dimensions, and regional disparities in scholarly collaboration suggest systemic imbalances in the field. These findings underscore the urgent need for future research to broaden linguistic and geographic inclusivity, prioritize pragmatic and pedagogical applications, and promote equitable international partnerships.

Nevertheless, this study is not without limitations. The exclusive reliance on the Scopus database may have led to the exclusion of relevant studies published in non-indexed regional journals or non-English outlets, potentially underrepresenting research from the Global South. In addition, the decision to include only peer-reviewed journal articles may omit practical reports or grey literature that could enrich pedagogical interpretations. Moreover, the analysis was conducted by a single coder, which, despite rigorous criteria, may introduce subjective interpretation or overlook nuances in data categorization.

Future research could address these limitations by triangulating multiple databases (e.g., Web of Science, ERIC, or Dimensions), incorporating multilingual sources, and expanding the scope to include grey literature or teacher-generated corpora. It is also recommended that future studies explore underexamined language pairings, especially involving African and indigenous languages, and give greater attention to contrastive pragmatics, including speech acts, politeness strategies, and interactional norms across cultures. In terms of pedagogical implications, the results encourage language educators to integrate corpus-informed materials that emphasize real-world discourse patterns, metaphorical usage, and contrastive lexical features. Teacher training programs can also benefit from exposure to contrastive corpus tools and methods to foster critical awareness of language variation and cross-cultural meaning.

REFERENCES

- Aijmer, K. (2020). Contrastive Pragmatics and Corpora. *Contrastive Pragmatics*, 1(1), 28–57. <https://doi.org/10.1163/26660393-12340004>
- Al-Khaza'leh, B. A., & Alzubi, A. A. F. (2022). The Metaphorical Conceptualization of Love in English and Arabic Songs: A Contrastive Study. *Theory and Practice in Language Studies*, 12(6), 1189–1199. <https://doi.org/10.17507/tpls.1206.19>
- Altameemi, Y. M. (2024). State-of-the-Art Review of the Corpus Linguistics Field From the Beginning Until the Development of ChatGPT. *Theory and Practice in Language Studies*, 14(2), 423–431. <https://doi.org/10.17507/tpls.1402.13>
- Andersen, G. (2023). Contrastive Corpus Studies of Pragmatic Markers. *Contrastive*

- Pragmatics*, 4(2), 173–177. <https://doi.org/10.1163/26660393-bja10090>
- Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, 30(2), 141–161. <https://doi.org/10.17250/KHISLI.30.2.201308.001>
- Arroyo, B. L., & Roberts, R. P. (2017). Genre and register in comparable corpora: An English/Spanish contrastive analysis. *Meta*, 62(1), 114–136. <https://doi.org/10.7202/1040469ar>
- Auni, L., & Manan, A. (2023). A Contrastive Analysis of Morphological and Syntactic Aspects of English and Indonesian Adjectives. *Studies in English Language and Education*, 10(1), 403–423. <https://doi.org/10.24815/siele.v10i1.27401>
- Blagoeva, D., Jaskot, M. P., & Sosnowski, W. (2020). A lexicographical approach to the contrastive analysis of Bulgarian and Polish phraseology. *Cognitive Studies*, 19(19), 1–17. <https://doi.org/10.11649/cs.1923>
- Charteris-Black, J. (2004). *Critical Metaphor Analysis BT - Corpus Approaches to Critical Metaphor Analysis*. London: Palgrave Macmillan UK. https://doi.org/10.1057/9780230000612_12
- Cheng, C. (2014). A contrastive study of English and Chinese book reviews on linguistics: Perspective of attitudinal meanings. *Theory and Practice in Language Studies*, 4(5), 1009–1016. <https://doi.org/10.4304/tpls.4.5.1009-1016>
- Culpeper, J., & Haugh, M. (2014). Pragmatics and the English Language. In *Pragmatics and the English Language*. London: Palgrave Macmillan UK. <https://doi.org/10.1007/978-1-137-39391-3>
- Curry, N. (2023). Question illocutionary force indicating devices in academic writing A corpus-pragmatic and contrastive approach to identifying and analysing direct and indirect questions in English, French, and Spanish. *International Journal of Corpus Linguistics*, 28(1), 91–119. <https://doi.org/10.1075/ijcl.20065.cur>
- Curry, N., & Chambers, A. (2017). Questions in English and French Research Articles in Linguistics: A Corpus-Based Contrastive Analysis. *Corpus Pragmatics*, 1(4), 327–350. <https://doi.org/10.1007/s41701-017-0012-0>
- De Cesare, A.-M., & Garassino, D. (2015). On the status of exhaustiveness in cleft sentences: An empirical and cross-linguistic study of English also-/only-clefts and Italian anche-/solo-clefts. *Folia Linguistica*, 49(1), 1–56. <https://doi.org/10.1515/flin-2015-0001>
- Dobrić Basanež, K. (2024). A Corpus-Based Contrastive Analysis of Metaphorical Collocations of the Terms Right and Pravo: Do Different Legal Cultures Give Rise to Different Metaphors? *ELOPE: English Language Overseas Perspectives and Enquiries*, 21(1), 169–183. <https://doi.org/10.4312/elope.21.1.169-183>
- Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133(March), 285–296. <https://doi.org/10.1016/j.jbusres.2021.04.070>
- Elgobshaw, A. E. (2024). Conceptualization of Morphological Roots in Arabic and English: A Contrastive Analysis. *World Journal of English Language*, 14(5), 436–443. <https://doi.org/10.5430/wjel.v14n5p436>
- Elhami, A. (2020). A socio-pragmatic perspective of spanish and persian greeting. *Theory and Practice in Language Studies*, 10(9), 1009–1014. <https://doi.org/10.17507/tpls.1009.01>
- Elorza, I. (2014). Newsworthiness, attribution and lexicogrammatical strategies in two types of news articles in English and Spanish. *Topics in Linguistics*, 14(1), 16–33. <https://doi.org/10.2478/topling-2014-0009>
- García, N. R. (2021). Contrastive Linguistics and Translation Studies Interconnected: The Corpus-based Approach. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 1(1), 393–406. <https://doi.org/10.52034/LANSTTS.V1I1.27>
- Granger, S., Gilquin, G., & Meunier, F. (2015). *The Cambridge Handbook of Learner Corpus*

- Research. Cambridge: Cambridge University Press.
<https://doi.org/10.1017/CBO9781139649414>
- Hasselgård, H. (2020). Corpus-based contrastive studies: Beginnings, developments and directions. *Languages in Contrast*, 20(2), 184–208.
<https://doi.org/10.1075/lic.00015.has>
- Haugh, M. (2013). Im/politeness, social practice and the participation order. *Journal of Pragmatics*, 58, 52–72. <https://doi.org/10.1016/j.pragma.2013.07.003>
- Hebedová, P. (2024). Deictic motion verbs and anchoring the direction of motion in Estonian, Finnish and Czech. *Lahivordlusi Lahivertailuja*, 2024(34), 15–48.
<https://doi.org/10.5128/LV34.01>
- Hong, N. B., & Pham, H. (2024). A Contrastive Analysis of Rhetorical Structures of English and Vietnamese Linguistic Research Articles. *3L: Language, Linguistics, Literature*, 30(1), 191–207. <https://doi.org/10.17576/3L-2024-3001-14>
- Hyland, K. (2016). Academic publishing and the myth of linguistic injustice. *Journal of Second Language Writing*, 31, 58–69. <https://doi.org/10.1016/j.jslw.2016.01.005>
- Ibrahim, R. K. (2015). *Translation oriented corpus-based contrastive linguistics*. 61(3), 381–393. <https://doi.org/https://doi.org/10.1075/babel.61.3.04kha>
- Izquierdo, M., & Pérez Blanco, M. (2023). Interactional metadiscourse: Building rapport and solidarity in informational-persuasive discourse. An English-Spanish case study. *Journal of Pragmatics*, 216, 106–120. <https://doi.org/10.1016/j.pragma.2023.08.005>
- Jafari, S. S. (2014). A cross-linguistic study of english and persian prepositions. *Journal of Language Teaching and Research*, 5(3), 689–697. <https://doi.org/10.4304/jltr.5.3.689-697>
- Jenkins, J. (2013). *English as a Lingua Franca in The International University: The Politics of Academic English Language Policy*. London: Routledge.
<https://doi.org/10.4324/9780203798157>
- Juodintytė-Kuznetsova, K. (2015). Contrastive Semantic Analysis of Contract Law Terms in Studies of Foreign Language for Specific Purposes. *Sustainable Multilingualism*, 7(7), 58–72. <https://doi.org/10.7220/2335-2027.7.3>
- Kerimbayeva, K., & Beisenbai, A. (2024). Wealth/Poverty Opposition in English and Kazakh: A Comparative Study. *World Journal of English Language*, 14(5), 596–604.
<https://doi.org/10.5430/wjel.v14n5p596>
- Kotorova, E. (2021). Methodology of cross-cultural description of speech behaviour: A case study of invitation in German and Russian discourse. *GEMA Online Journal of Language Studies*, 21(1), 1–19. <https://doi.org/10.17576/gema-2021-2101-01>
- Kövecses, Z. (2002). *Metaphor: A Practical Introduction*. New York: Oxford University press.
- Kralj, S., & Polajnar, J. (2021). Dynamics of Discursive Themes in Digital Discourse -a Contrastive Discourse-Linguistic Analysis of Slovene and German Commentary Forums. *Linguistica (Slovenia)*, 61(1), 37–65.
<https://doi.org/10.4312/linguistica.61.1.37-65>
- Lei, L., & Liu, D. (2019). Research trends in applied linguistics from 2005 to 2016: A bibliometric analysis and its implications. *Applied Linguistics*, 40(3), 540–561.
<https://doi.org/10.1093/applin/amy003>
- Leuschner, T., & Van Den Nest, D. (2015). Asynchronous grammaticalization: V1-conditionals in present-day English and German. *Languages in Contrast*, 15(1), 34–64.
<https://doi.org/10.1075/lic.15.1.03leu>
- Levin, M., & Herold, J. S. (2024). English complex premodifiers and their German and Swedish correspondences: The case of hyphenated premodifiers in a non-fiction corpus. *Languages in Contrast*, 24(1), 5–32. <https://doi.org/10.1075/lic.00033.lev>

- Lin, Z., & Lei, L. (2020). The research trends of multilingualism in applied linguistics and education (2000-2019): A bibliometric analysis. *Sustainability (Switzerland)*, 12(15), 1–14. <https://doi.org/10.3390/su12156058>
- Lyu, S., & Wang, Y. (2015). Karin Aijmer & Bengt Altenberg (eds.), Advances in corpus-based contrastive linguistics (Studies in Corpus Linguistics 54). *Journal of Linguistics*, 51(3), 677–682. <https://doi.org/10.1017/s0022226715000250>
- McKinley, J. (2019). Introduction: Theorizing research methods in the ‘golden age’ of applied linguistics research. *The Routledge Handbook of Research Methods in Applied Linguistics*, 1(1), 1–12. <https://doi.org/10.4324/9780367824471-1>
- Mohd Jalis, F. M., & Abd Rahim, N. (2014). Contrastive analysis of German and Malay modal verbs. *Asian Social Science*, 10(6), 81–89. <https://doi.org/10.5539/ass.v10n6p81>
- Moinzadeh, A., Dezhara, S., & Rezaei, O. (2012). A Contrastive Study of L1 and L2 Acquisition. *Theory and Practice in Language Studies*, 2(1), 66–70. <https://doi.org/10.4304/tpls.2.1.66-70>
- Pablos-Ortega, C. de. (2015). Thank you for a lovely day! Contrastive Thanking in Textbooks for Teaching English and Spanish as Foreign Languages . *Pragmática Sociocultural / Sociocultural Pragmatics*, 3(2), 150–173. <https://doi.org/10.1515/soprag-2015-0023>
- Page, M. J., McKenzie, J. E., Bossuyt, P., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *Medicina Fluminensis*, 57(4), 444–465. https://doi.org/10.21860/medflum2021_264903
- Qin, X. (2015). Recent developments in corpus linguistics and corpus-based research / Department of Linguistics and Modern Language Studies at the Hong Kong Institute of Education. *Language Teaching*, 48(1), 156–160. <https://doi.org/10.1017/S0261444814000330>
- Römer, U., & Wulff, S. (2010). Applying corpus methods to written academic texts: Explorations of MICUSP. *Journal of Writing Research*, 2(2), 99–127. <https://doi.org/https://doi.org/10.17239/jowr-2010.02.02.2>
- Ruzaitė, J. (2024). Disinformation Crossing Spaces and Language Borders: A Contrastive Analysis of English and Lithuanian. *NJES Nordic Journal of English Studies*, 23(2), 268–298. <https://doi.org/10.35360/njes.v23i2.39193>
- Sareh, N. (2023). Comparison of the Structure of Words in Patani Malay and Thai. *Issues in Language Studies*, 12(2), 1–17. <https://doi.org/10.33736/ils.5552.2023>
- Sejdiu-Rugova, L., & Krijezi, M. (2018). The current use of the past imperfect tense in Albanian and Serbian - A case study. *Slavia Meridionalis*, 18(1487), 1–23. <https://doi.org/10.11649/sm.1487>
- Šimůnková, R. (2018). English present participles and gerunds functioning as adverbials and their translations into Czech. *Discourse and Interaction*, 11(1), 67–85. <https://doi.org/10.5817/DI2018-1-67>
- Skjott Linneberg, M., & Korsgaard, S. (2019). Coding qualitative data: a synthesis guiding the novice. *Qualitative Research Journal*, 19(3), 259–270. <https://doi.org/10.1108/QRJ-12-2018-0012>
- Sosnowski, W. P., & Satoła-Staskowiak, J. (2019). A contrastive analysis of feminines in Bulgarian, Polish and Russian. *Cognitive Studies*, 19(1922), 1–12. <https://doi.org/10.11649/cs.1922>
- Wu, C., & Baccanello, J. (2019). Using Contrastive Terminology Analysis in Teaching a Foreign Language. *Eurasian Journal of Applied Linguistics*, 5(3), 461–471. <https://doi.org/10.32601/ejal.651339>