CSCanada
Studies in Literature and Language
Vol. 31, No. 1, 2025, pp. 25-32
DOI:10.3968/13857

ISSN 1923-1555[Print]
ISSN 1923-1563[Online]
www.cscanada.net
www.cscanada.org

Corpus-driven Research on Foreign French Teaching

YANG Tong[a],*; FAN Xinyan[a]

[a] School of Foreign Languages, North China Electric Power University, Beijing, China.
* Corresponding author.

Received 21 June 2025; accepted 30 July 2025
Published online 26 August 2025

Abstract

As an important methodology in modern language teaching, the corpus-based approach offers a core advantage: it provides large-scale language samples embedded in authentic contexts for language learning. This study systematically explores the framework of corpus linguistics from two dimensions: theoretical construction and teaching practice. First, it outlines the fundamental theoretical framework of corpus linguistics and its historical development. Then, adopting an empirical research approach, the study takes the French polysemous word "éolien" as the focus of its research. Through detailed corpus analysis, it reveals the systematic characteristics of the word's semantic network and its applicability in teaching. At the teaching practice level, the study designs a three-stage inquiry-based learning model of "observation-categorization-induction." This research provides a practical theoretical framework and paradigm for corpus-driven foreign language teaching.

Key words: Corpus teaching; Corpus linguistics; Foreign French teaching; Corpus driven; Index

Yang, T., & Fan, X. Y. (2025). Corpus-driven Research on Foreign French Teaching. *Studies in Literature and Language, 31*(1), 25-32. Available from: http://www.cscanada.net/index.php/sll/article/view/13857
DOI: http://dx.doi.org/10.3968/13857

1. INTRODUCTION

The integration of large-scale and diverse corpus resources into corpus-based instruction effectively overcomes the constraints inherent in traditional language teaching, which has long depended on manually curated materials and abstract grammatical rules. This shift paves the way for a more empirical and data-driven approach to pedagogy. The theoretical contributions of prominent linguists such as J. R. Firth, Sinclair, and Halliday have provided a solid foundation for corpus-based language instruction. Their insights lend substantive support to its development in key areas including semantic analysis, the elucidation of linguistic patterns, and functional interpretation. Corpus-based language instruction proves particularly effective in the domains of vocabulary acquisition, grammatical competence, and reading comprehension, markedly enhancing learners' ability to understand and apply language in context.

Despite these advances, current academic discourse reveals room for further development in the systemic study of corpus pedagogy. A more in-depth exploration of its theoretical underpinnings, historical trajectory, and practical implementation is crucial for fostering innovation in language education and meeting the evolving demands of contemporary society.

2. THE FUNDAMENTAL CONCEPTS AND THEORETICAL FOUNDATIONS OF CORPUS-BASED LANGUAGE TEACHING

2.1 Concept and Characteristics

Corpus-based language teaching refers to a pedagogical approach that integrates large-scale collections of authentic language texts into instructional practice. These corpora encompass a wide variety of language materials—including written texts, spoken dialogues, news reports, and literary works—that collectively represent diverse manifestations of language use in real-world contexts.

According to He Anping (2019), corpus-based language teaching is distinguished by two principal characteristics: (1) An emphasis on the frequency of linguistic usage; and (2) a focus on the contextual environment in which language occurs.

Corpus linguistics is grounded in authentic data of language usage and advocates the revelation of linguistic patterns through large-scale corpus analysis. This methodological approach directly provides the theoretical foundation for corpus-based teaching. Johns (1991) emphasizes that language constitutes a "usage-based system," rather than an abstract set of rules, and supports the induction of linguistic principles through observation of real language examples—commonly referred to as Data-Driven Learning. The core tools of corpus-based pedagogy, such as concordance lines in context, frequency analysis, and collocation patterns, are all derived from research methodologies established within corpus linguistics.

2.2 Theoretical Foundations

J. R. Firth's (1957) theory of contextual co-occurrence laid the semantic foundation for corpus-based language teaching. Its core proposition—"You shall know a word by the company it keeps"—emphasizes the importance of linguistic environment in understanding word meaning. In practice, this theory guides students to analyze contextual co-occurrence patterns within corpora to grasp semantic content. The principal value lies in providing a theoretical basis for context-driven semantic analysis, thereby enabling learners to accurately comprehend lexical meanings.

Building upon Firth's work, Sinclair (1991) introduced the theory of lexical behavior, which underscores the probabilistic nature of language use over absolute rule-governed structures. The vast authentic language data found in corpora reveals probabilistic information such as collocations and frequency patterns, furnishing corpus-based pedagogy with a framework that closely reflects real-world language use and assists both instructors and learners in recognizing the actual regularities of language usage.

Swales (1990) defines genre as a type of language event with specific communicative purposes, characterized by distinctive schematic structures and linguistic features. Through genre analysis of texts in corpora, one can uncover the conventions of various genres. This theory links genre with pedagogy, promoting a deep integration between corpora and teaching, and aiding learners in mastering linguistic norms across genres to adapt to diverse communicative contexts. Halliday's (1994) systemic functional linguistics proposes that language has three metafunctions—ideational, interpersonal, and textual—and operates as a systemic network interconnected with context. Authentic data from corpora offer rich cases for applying this theory. Its theoretical value lies in providing a functional framework for corpus-based teaching, enabling teachers and students to analyze how language functions in different contexts, thereby enhancing communicative competence.

With the advancement of natural language processing technologies, corpus tools have evolved from simple retrieval to deep analysis. Software such as AntConc and WordSmith support part-of-speech tagging, collocational strength calculations, and visualized presentation. Lev Vygotsky's (1978) social constructivist theory has shifted corpus-based teaching from a "teacher-centered" to a "student-centered" approach. In Ellis's (1985) theory of second language acquisition, learners analyze and internalize the frequency distribution patterns of input, gradually mastering a wide range of language structures and abstracting frequency-oriented linguistic rules, thereby achieving a transformation of language knowledge from explicit to implicit. Sweller's (1988) cognitive load theory offers psychological grounding for corpus-based instruction; in such contexts, large volumes of unfiltered or unprocessed authentic data may increase learners' intrinsic cognitive load.

In recent years, corpus-based teaching has demonstrated a trend toward interdisciplinary integration. Ecological linguistics emphasizes the dynamic interaction between language and the social environment. Fairclough's (1995) critical discourse analysis (CDA) introduces a dimension of social critique into corpus pedagogy.

3. DEVELOPMENT OF CORPUS-BASED TEACHING

3.1 Early Exploratory Stage

In the 1940s, linguists such as Fries & Traver (1950) were among the first to introduce corpus data into research on foreign language teaching methods. At the time, structuralist linguistics prevailed, emphasizing detailed descriptions of language structures. Structuralist scholars sought to analyze linguistic structures with greater precision using corpus data, thereby refining foreign language pedagogy. For instance, by examining the frequency and contextual distribution of vocabulary and grammatical structures in large samples of spoken or written texts, they aimed to identify pedagogical priorities. However, corpus linguistics was still in its infancy, and computer technology had not yet become widespread. Corpus collection, organization, and analysis were primarily conducted manually, resulting in low efficiency and limited scale. Mainstream academia continued to focus on rule-based linguistic research, and corpus linguistics had yet to gain broad recognition. Related instructional applications were confined to pioneering efforts, without forming a systematic teaching model or achieving widespread implementation.

3.2 Progressive Development Stage

One of the earliest international efforts to apply corpora in language teaching emerged in the early 1990s, when Johns pioneered the concept of "Data-Driven Learning". Specifically, this approach involves using corpus concordancing tools to present lines of target words, guiding students to observe collocates surrounding the target items and infer patterns of lexical and grammatical co-selection. This method addresses challenges in learning confusing or polysemous words.

Domestically, an early adopter of corpus-assisted English teaching was the corpus team at the School of Foreign Languages, South China Normal University. Since 1998, they have offered undergraduate courses in corpus-assisted English pedagogy for pre-service teachers, which gradually expanded to postgraduate courses and English teacher training programs at the elementary, secondary, and tertiary levels.

3.3 Widespread Application Stage

Since the beginning of the 21st century, corpus-based approaches to language teaching have gained increasing attention. This methodology relies on authentic corpus data to inform instruction, extracting patterns from extensive real-world language use to cultivate learners' awareness of autonomous language learning and practical communicative ability. The core idea is to treat the corpus as a primary source of language input, guiding learners to locate, observe, generalize, and apply linguistic knowledge through analysis of features in authentic contexts.

Research has shown that corpora offer notable advantages across multiple aspects of pedagogy. In vocabulary learning, corpora present numerous instances of word usage in genuine contexts, helping students grasp subtle semantic distinctions and typical collocational patterns. By providing authentic and natural language samples, students gain exposure to idiomatic expressions and enhanced intuitive understanding of language. In instructional design, teachers can leverage corpus data to assess learners' language proficiency and needs, thereby crafting targeted activities such as role plays or text analyses to stimulate learner engagement. Sinclair's *Reading Concordances* (2003) exemplifies this approach by guiding students from basic to complex understanding of vocabulary and grammar through step-by-step questioning based on abundant examples, fostering their independent learning abilities.

4. APPLICATION OF CORPORA IN LANGUAGE TEACHING

Corpora are typically employed in language instruction through two pedagogical approaches: the inductive approach and the deductive approach, also known as inductive reasoning and deductive reasoning. Inductive reasoning refers to an argumentation process that moves from particular instances to generalizations. It requires the collection of a substantial amount of linguistic data, from which universal patterns and grammatical rules are derived. This method has gained widespread adoption in foreign language teaching abroad. In contrast, the deductive approach—more commonly used in domestic settings—introduces language rules directly to learners, guiding them to apply these rules in practice. Corpus-based pedagogy integrates both approaches: learners generalize linguistic patterns through inductive reasoning and subsequently apply these self-derived rules in actual usage. The corpus-driven approach represents a key methodology in corpus linguistics, emphasizing bottom-up analysis drawn solely from the corpus itself. Through exhaustive observation, categorization, and analysis of all linguistic instances in a corpus, researchers inductively formulate hypotheses or even theoretical conclusions about language use. This approach aims to uncover patterns, regularities, and features inherent in authentic language data. In foreign language education, corpus-driven teaching is particularly effective in areas such as vocabulary and grammar instruction.

4.1 Vocabulary Instruction

In French language teaching, instructors can utilize corpus analysis tools such as AntConc to input the word "éolien" and retrieve its collocates across different contexts—for example, "le parc éolien" (wind farm), "l'énergie éolienne" (wind energy), and "l'éolien terrestre" (onshore wind power). These collocations reflect varying semantic emphases. Through comparative analysis of authentic corpus data, students can intuitively grasp the dynamic usage patterns of "éolien". For synonymous terms like "bon" and "excellent", the concept of semantic prosody effectively distinguishes their nuances. By searching the French National Corpus (CNRTL), it is observed that bon tends to occur in everyday contexts (e.g., "bonhomme", "bonne idée"), while excellent frequently co-occurs with evaluative nouns (e.g., "excellent résultat", "excellent service"), clearly exhibiting a positive semantic prosody.
Teachers may employ corpus tools to contrast the usage of "prendre" and "recevoir". Based on search results from the French National Corpus: typical collocations of "prendre" include "prendre une décision" (make a decision), "prendre le déjeuner" (have lunch), and "prendre le train" (take the train); whereas "recevoir" commonly collocates with expressions such as "recevoir un cadeau" (receive a gift), "recevoir des invités" (host guests), and "recevoir une formation" (receive training). The comparison reveals that "prendre" emphasizes active acquisition, while "recevoir" denotes passive reception. Through extensive analysis of authentic corpus data, students can clearly understand the distinction between these two verbs.

4.2 Grammar Instruction

Corpora can reveal how grammatical structures are used in authentic contexts. Teachers may extract dozens of sentences featuring the target grammar from corpora and, through a hierarchy of guiding questions, gradually lead students to independently discover and internalize grammatical knowledge.

Teachers can also use corpora to analyze students' grammatical errors. For example, by collecting common mistakes found in student compositions—such as confusion between the verbs "savoir" and "connaître"—these can be input into a corpus for comparative search. If a student writes "Je connais que le soleil se lève à l'est," searching for "Je connais que" and "Je sais que" in the corpus reveals that "Je sais que" accounts for 91% of occurrences in contexts expressing knowledge of facts or information. Frequent collocations also include "sais + infinitive", as in "Je sais nager" ("I can swim"). In contrast, "Je connais que" is virtually nonexistent ; connaître is more commonly used to express familiarity with people or places, such as "Je connais Paris" ("I know Paris") or "Je connais Paul" ("I know Paul"). This kind of corpus-based contrastive analysis helps students understand lexical distinctions through authentic contexts and avoid misusage caused by similar Chinese definitions, thereby establishing accurate cognition from verb to context to meaning.

5. CORPUS-DRIVEN RESEARCH ON THE APPLICATION OF FRENCH AS A FOREIGN LANGUAGE TEACHING

In traditional teaching models, the teacher primarily serves as the transmitter of knowledge, while learners occupy a relatively passive "receiving" position, unidirectionally absorbing the "learning content" provided by the instructor as the "knowledge provider". In contrast, corpus-driven research on teaching French as a foreign language presents a different pedagogical picture : the teacher becomes a learning facilitator, guiding learners to actively explore the internal rules of the language through carefully designed instructional activities. Within this innovative teaching model, learners participate as inquirers in the learning process, conducting linguistic discovery under expert guidance; the teacher, by providing scientifically selected sequences of corpus data, fully assumes a dual role as both "learning facilitator" and "cognitive mediator" (Kübler, 2014: 44). It is particularly important to emphasize that this approach imposes strict requirements on the complexity of corpus material. The selected linguistic content should follow a progressive sequencing principle to avoid compromising learners' comprehension due to inappropriate material complexity. This study utilizes a 1.08-million-word French corpus from the electricity sector. As a researcher, the teacher employs AntConc software to conduct linguistic analyses—syntactic, lexical, and semantic—using the word "éolien" as an example, alongside recommended teaching activities.

5.1 Linguistic Analysis

5.1.1 Syntactic Analysis

The word "éolien" belongs to two grammatical categories : noun and adjective. As a noun, it can be preceded by adjectives and nouns, and it may also be followed by an adjective. As an adjective, it can modify a noun or appear after a noun or another adjective.

To enhance clarity, Table 1 presents the most frequently occurring words in each grammatical category, derived from data extracted from the electricity domain corpus.

Table 1
Frequency of Preceding and Following Word Classes Associated with Two Forms of *éolien*(Noun vs. Adjective)

	avant	pivot	après
préposition	en (2), avec (1), sans (1), entre (1)		à (71), en (34), de (4), dans (26)
préfixe	micro- (1)		
verbe		éolien (nom)	commencer (2), constituer (1), être (1)
adjectif			offshore (15), terrestre (14)
nom	le parc (112), l'énergie(76), l'électeicité (22), le gisement (20), la capacité (18), le système (13), le projet (13), la ferme (7), la centrale (2)	éolien (adjectif)	
adjectif	spatial (32), grand (7), petit (5)		industriel (1), précédent (1)

As shown in Table 1, "éolien" most frequently occurs as an adjective in this corpus, with nouns appearing at the highest frequency accordingly. Furthermore, we observe that "éolien" is often used in conjunction with prepositions such as "à" and "en", which are placed before or after "éolien" to clarify its functional usage.

5.1.2 Lexical Analysis

In vocabulary instruction, corpora play a significant role in revealing lexical collocations and semantic prosody. Based on descending frequency of co-occurrence, éolien as an adjective forms specific expressions with words such as "parc", "énergie", and "gisement". As a noun, it forms specific expressions with words like "terrestre" and "offshore". Whether these specific expressions constitute fixed phrases requires further analysis.

G. Gross's (1988) theory of nine syntactic transformations forms part of his lexical grammar framework, which serves to analyze the structural

properties of linguistic expressions and to determine whether they are free expressions (expression libre), fixed expressions (expression figée), or semi-fixed expressions (expression semi-figée). The nine types of transformations include : predicativity (prédicativité), nominalization (nominalisation), selection restriction (restriction de sélection), number variation (variation en nombre), adverb addition (adjonction d'un adverbe), adjective addition (adjonction d'un adjectif), determiner addition (adjonction d'un déterminant), word order transposition (transposition), and component omission (suppression). In this analytical framework, a "+" typically indicates that the transformation is both theoretically and empirically feasible, yielding expressions that are grammatically and semantically acceptable. Conversely, a "-" signifies that the transformation is not viable, producing forms that are either linguistically unnatural or semantically incoherent.

Using the phrase "éolien offshore" as an example, we categorize and analyze its expression type based on the nine syntactic transformations proposed by Gross (1988). The results are presented in Table 2.

Table 2
Expression Type Classification of éolien offshore Based on Transformation Criteria

Prédicativité : éolien est offshore	—
Nominalisation : offshore de l'éolien	—
Restriction de sélection : éolien terrestre	+
Variation en nombre : éoliens offshores	—
Adjonction d'un adverbe : éolien très offshore	—
Adjonction d'un adjectif : éolien offshore et terrestre	+
Effacement d'un adjectif : éolien	+
Figement du premier terme : paiement offshore	+
Équivalence: adjectif=de nom : éolien de offshore	—
Éolien offshore	Expression semi-figée

According to Gross's (1988) typology of nine syntactic transformations, the phrase éolien offshore permits four while resisting five, qualifying it as a semi-fixed expression. Further analysis of frequent collocations with éolien—including parc, énergie, gisement, and terrestre—reveals that these combinations also exhibit semi-fixed characteristics. This indicates a lexical behavior that balances flexibility and structural convention, enriching expressive depth and enhancing the dimensionality of language use.

5.1.3 Semantic Analysis

Table 3 presents twenty randomly selected indexed entries of "éolien" in varying forms and syntactic functions, drawn from a specialized corpus on electricity. These entries serve to illustrate the semantic range of éolien across different contexts, enabling a detailed interpretation of its lexical behavior in technical discourse.

Table 3
Semantic Categorization of éolien Sentence Variants Extracted from the Corpus

	Contexte gauche	pivot	Contexte droit
1	Comme pour l'	éolien	, la troisième rang de la France au niveau
2	sa stratégie de developpement basé sur l'	éolien	et le PV, c'est le « mix-renouvelable » de
3	La puissance installée par le parc	eolien	et la surface des panneaux photovoltaîque
4	Toutefois, les	éoliennes	à axe vertical présentent plusieurs inconvénients
5	D'après certaines études, si les	éoliennes	volantes captaient 1% des vents
6	Le système de conversion d'énergie	éolienne	est composé de deux convertisseurs
7	Cette incrémentation de la part d'énergie	éolienne	va en particulier s'accélérer
8	Concernant la commande des grandes	éoliennes	, nous souhaitons en particulier continuer
9	Les commandes d'conçues	éolienne	pendant mes années
10	laboratoire de tests d'	éoliennes	au monde nous fait espérer
11	L'association d'	éoliennes	à des systèmes de stockage
12	la connexion de parcs	éoliens	offshores au réseau continental.
13	la situation ou système	éolien	est couplé directement au réseau
14	Energie renouvelable, Parc	éolien	, stockage d'énergie,
15	pour qu'une	éolienne	commence à produire de l'électricité.
16	pour mieux intégrer les productions	éoliennes	et solaires.
17	Il ressort que le gisement	éolien	est plus élevé le long
18	La ressource	éolienne	dans plusieurs pays
19	L'excédent de la production de l'	éolienne	est utilisé pour produire l'hydrogène.
20	Si la production du parc	éolien	dépasse la demande du réseau électrique

The twenty indexed sentences in Table 3 are first grouped by grammatical category: sentences 1, 2, 4, 5, 8, 9, 10, 11, 15, and 19 form noun phrases, while sentences 3, 6, 7, 12, 13, 14, 16, 17, 18, and 20 constitute adjectival phrases. Within the noun phrase group, sentences 1 and 2 convey the meaning of "wind technology" , whereas sentences 4, 5, 8, 9, 10, 11, 15, and 19 align with the semantic field of "wind turbine". For the adjectival phrases, sentences 3, 12, 13, 14, 17, and 20 pertain to the meaning "wind-related", while sentences 6, 7, 16, and 18 correspond to "wind-powered" or "pertaining to wind energy".

Within this corpus, the lexical item "éolien" manifests four distinct meanings. Both its masculine ("éolien") and feminine ("éolienne") forms can function nominally, yet they denote different concepts : "éolien", as a neologism,

refers to the entirety of technologies associated with wind energy and its industrial applications, whereas “éolienne” designates the apparatus that transforms wind’s kinetic energy into mechanical or electrical power. As an adjective, “éolien” possesses two principal senses: one relates to wind in a general atmospheric context, while the other pertains specifically to wind turbines. Illustrative examples include “centrale éolienne” (wind power station), “parc éolien” and “ferme éolienne” (both referring to wind farms where multiple éoliennes operate collectively to generate electricity). The expression “énergie éolienne” (wind energy) aligns with the more general term “énergie du vent” (energy derived from wind). In most corpus instances, “éolien” is employed adjectivally, with “parc éolien” and “énergie éolienne” emerging as its most frequent collocational patterns within the domain of energy discourse.

Based on syntactic analysis, “éolien” exhibits two grammatical categories : adjective and noun. The most frequent co-occurrences involve nouns and prepositions. From a lexical perspective, the majority of collocations formed with “éolien” are semi-fixed expressions, combining structural flexibility with conventional usage. This hybrid nature contributes to nuanced layers of linguistic expression. Semantically, “éolien” carries four distinct meanings, attesting to its polysemous nature. Among these analytical dimensions, the semantic analysis proves especially pedagogically valuable. In the context of French language instruction, corpus-based examination of “éolien” offers learners concrete instances to better grasp its multiple senses.

5.2 Recommended Corpus-Based Pedagogical Activities in Linguistics

This study adopts a corpus-driven research methodology, using the term “éolien” as a case study to develop instructional activities via the indexing tool AntConc 4.3.1. Educators may extract occurrences of “éolien” and its surrounding context from the corpus or apply the wildcard search technique (éolien*) to retrieve all indexed variants of the term across its morphological forms. The instructional sequence begins with teachers selecting and refining representative concordance lines, manually excluding complex or potentially misleading instances to facilitate learners’ comprehension of unfamiliar vocabulary. The curated examples are then distributed to students, who engage in observation, classification, and inductive synthesis of usage patterns.

5.2.1 Observation and Classification

Learners are guided to sequentially examine each example presented in Table 3 and attempt to infer the meaning of “éolien” and its various morphological forms without consulting a dictionary. Given their lack of prior exposure to this term, learners tend to categorize the examples based on the most salient grammatical features—specifically distinguishing between nominal and adjectival usages—as demonstrated in Table 4.

Table 4
Preliminary Categorization of Examples by Learners

Grammatical Category	Example Numbers
Noun	1、2、4、5、8、9、10、11、15、19
Adjective	3、6、7、12、13、14、16、17、18、20

Building upon the initial categorization, learners may further investigate the semantic properties of the noun forms. Teachers instruct learners to distinguish between masculine and feminine noun examples, resulting in the distribution presented in Table 5.

Table 5
Gender-Based Classification of the Noun Forms of éolien

Masculin noun	Comme pour l’éolien, la troisième rang de la France au niveau communautaire traduit mal son retard
	sa stratégie de developpement basé sur l’éolien et le PV, c’est le « mix- renouvelable » de
Feminin noun	Toutefois, les éoliennes à axe vertical présentent plusieurs inconvénients
	D’après certaines études, si les éoliennes volantes captaient 1% des vents
	Concernant la commande des grandes éoliennes , nous souhaitons en particulier continuer
	Les commandes d’conçues éolienne pendant mes années
	laboratoire de tests d’ éoliennes au monde nous fait espérer
	l’association d’ éoliennes à des systèmes de stockage
	pour qu’une éolienne commence à produire de l’électricité.
	L’excédent de la production de l’éolienne est utilisé pour produire l’hydrogène.

An examination of Table 5 indicates that éolien refers to “wind power technology”, while éolienne denotes a “wind turbine”.

To distinguish the adjectival usage of “éolien”, instructors may select pertinent corpus-based examples and distribute them to learners. Students are tasked with identifying and annotating occurrences of “éolien” and its morphological variants, focusing specifically on the semantic and syntactic characteristics of the nouns they modify.

En revanche, la filière de production éolienne en milieu terrestre présente des menaces que les industriels et investisseurs prennent en compte.

Par la suite, nous allons intégrer un générateur éolien dans le réseau.

Dans les pays qui accordent aujourd’hui une place importante à l’énergie éolienne tels que l’Allemagne...

Ces parcs éoliens utilisent essentiellement (dans notre pays) des unités possédant les caractéristiques.

La filière éolienne offshore a connu un démarrage plus tardif, précisément en 1991.

Les systèmes éoliens transforment l’énergie cinétique

du vent en énergie électrique à travers des aérogénérateurs. The analysis reveals a semantic distinction between the feminine adjective "éolienne" and its masculine counterpart "éolien". The former tends to modify abstract nouns such as "production" (output), "énergie" (energy), and "filière" (sector), whereas the latter is more commonly associated with concrete entities, including "générateur" (generator), "parc" (power plant), and "système" (system). The phrase "l'énergie éolienne" is synonymous with "l'énergie du vent", positioning "éolienne" as a lexical representation of wind. In contrast, "éolien"—when used adjectivally—refers to tangible components involved in wind energy production, such as wind farms, turbines, and associated systems.

5.2.2 Generalization

Through the instructional activities described above, learners are generally able to infer the meanings of "éolien" independently. Teachers may divide learners into small groups to facilitate peer discussion and comparison of findings. Following these exchanges, each group may designate a representative to summarize their conclusions, after which the instructor may provide supplementary feedback and clarification. The semantic scope of "éolien" may then be systematically explained by the teacher, as illustrated in Table 6.

Table 6
Categorization of Example Usages

Keyword	Grammatical Role	Gender	Meaning
éolien	Noun	Masculine	All technologies associated with wind energy and its industry
		Feminine	A device that converts wind's kinetic energy into mechanical or electrical energy
	Adjective	Masculine	Relating to wind
		Feminine	Pertaining to wind turbines

Utilizing the corpus as an instructional tool, educators may guide learners through a gradual, layered analysis of linguistic data. This process enables learners to autonomously interpret and internalize the multiple semantic dimensions of vocabulary items. Given the corpus's ability to demonstrate grammatical structures in authentic usage contexts, corpus-based pedagogy can extend beyond vocabulary instruction to encompass grammar teaching, reading development, and broader language comprehension strategies.

6. CONCLUSION

This article offers a systematic examination of corpus-based pedagogy, underpinned by a diverse theoretical framework ranging from Firth's theory of collocational context to Lev Vygotsky's socioconstructivist paradigm. These perspectives collectively inform corpus-informed instruction through semantic analysis, linguistic pattern recognition, functional interpretation, and genre-based teaching. The evolution of corpus pedagogy has progressed from early manual exploration to computer-assisted development, and into its widespread application in the twenty-first century. Concepts such as data-driven learning and corpus-informed instructional design continue to shape and innovate this pedagogical approach. In classroom practice, corpora demonstrate significant advantages in both vocabulary and grammar instruction. They enable learners to uncover semantic prosody, acquire contextualized grammatical usage, and enhance reading efficiency—all through exposure to authentic language data.

Despite its pedagogical advantages, corpus-based instruction continues to face notable challenges concerning teacher proficiency, the quality of corpus resources, and student adaptability. Firstly, many instructors lack the technical expertise required to operate corpus tools effectively. Limited training opportunities in corpus query techniques hinder the integration of corpus-informed practices into mainstream language instruction. Secondly, the accessibility and quality of corpus resources remain problematic. Available corpora may be expensive, outdated, limited in scale, or inaccurately annotated. To meet specific teaching needs, instructors often resort to building their own small-scale corpora. However, this process demands considerable time and effort—a significant burden, particularly for those already overwhelmed by heavy teaching loads. Thirdly, corpus-based instruction demands learner autonomy, which may conflict with students' prior learning habits. Traditional classroom instruction typically favors direct explanation, with students playing a passive role. In contrast, corpus pedagogy requires learners to actively explore linguistic data. Some students may struggle to adapt, lacking both the motivation and analytical skills to engage with large volumes of authentic texts, thereby undermining the effectiveness of corpus-based teaching.

In the future, advances in theoretical research and technological development offer promising avenues for overcoming these constraints. With further innovation in interdisciplinary integration and pedagogical models, corpus-based instruction is poised to play a transformative role in the future of language education.

REFERENCES

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL, 17*(1), 47-64.

Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL, 19*(3), 307-328.

Ellis, R. (1985). *Understanding second language acquisition*. Oxford University Press.

Fairclough, N. (1995). *Critical discourse analysis: The critical study of language*. Longman.

Firth, J. R. (1957). A synopsis of linguistic theory. In *Studies in linguistic analysis*(pp. 10-32). Blackwell.

Flowerdew, L. (2008). *The pedagogic value of corpora: A critical evaluation*[Paper presentation]. Proceedings of the 8th Teaching and Language Corpora Conference, Lisbon, Portugal.

Fries, C. C., & Traver, A. A. (1950). *English word lists: A study of their adaptability for instruction*. George Wahr.

Gaston, G. (1988). Degré de figement de noms composés. *Langages, 23*(90), 57-72.

Halliday, M. A. K. (1994). *An introduction to functional grammar*(2nd ed.). Edward Arnold.

He, A. P. (2019). 何安平谈语料库与语言教学 [He Anping on corpus and language teaching]. *Corpus Linguistics Studies, 6*(1), 13-22.

He, A. P. (2022). 中小学英语教师语料库素养提升路径研究 [A study on the path to improve corpus literacy of primary and secondary school English teachers]. *Digital Teaching in Primary and Secondary Schools, 6*, 5-9.

Johns, T. F. (1991). Should you be persuaded: Two examples of data-driven learning materials. *English Language Research Journal, 4*, 1-16.

Kübler, N. (2014). Mettre en œuvre la linguistique de corpus à l'université - Vers une compétence utile pour l'enseignement/apprentissage des langues ? *Les Cahiers de l'Acedle, 11*(1), 37-77.

McCarthy, M. (2008). Accessing and interpreting corpus information in the teacher education context. *Language Teaching, 41*(4), 563-574.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, J. M. (2003). *Reading concordances: An introduction*. Longman.

Swales, J. M. (1990). *Genre analysis*. Cambridge University Press.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257-285.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.