JOURNAL OF LANGUAGE EDUCATION AND RESEARCH

Journal of Language Education and Research, 2025, 11(1), 284-304

Research Article

Will or *be going to*? Formulaic Patterns of Future Marking in L2 Academic English

Fatih Ünal Bozdağ *

ARTICLE INFO

Received:03.02.2025

Revised form: 13.03.2025

Accepted:22.03.2025

Doi: 10.31464/jlere.1631941

Keywords:

future time expression

learner English

cross-linguistic influence

formulaic language

temporal reference

ABSTRACT

This study examines how learners from different first language backgrounds express future time in academic English writing, focusing on *will* and *be going to* constructions. Drawing on a corpus of texts from 25 language backgrounds, the research reveals a strong preference for *will* (96.2%) over *be going to* (3.8%) across all groups. While *will* appears uniformly throughout texts, *be going to* is used more selectively, indicating distinct functional roles. The study documents variations in future expression density per text (2.30-4.85) and systematic differences in how future markers combine with different parts of speech, with Chinese learners showing particularly distinctive patterns. Cluster analysis reveals language family groupings, suggesting both universal constraints and L1 influences shape future expression. This finding shows *will* functions as a formulaic pattern in academic writing, advancing knowledge of how L2 learners develop temporal expressions.

Acknowledgments

Statement of Publication Ethics The current study does not require ethics committee approval.

Authors' Contribution Rate

Conflict of Interest None.

Reference Bozdağ, F., Ü. (2025). *Will* or *be going to*? Formulaic patterns of future marking in L2 academic English. *Journal of Language Education and Research, 11*(1), 284-304.

* Assist. Prof. Dr. ORCID ID: https://orcid.org/0000-0002-9959-4704, Osmaniye Korkut Ata University, English Translation and Interpretation, fatihbozdag@osmaniye.edu.tr

ISSN: 2149-5602

Introduction

The expression of future time in English presents a theoretical challenge at the intersection of temporal cognition and grammar. While many languages employ distinct morphological markers for future tense, English uses various constructions—modal auxiliaries, semi-auxiliaries, and temporal adverbials—that interact in specific ways to project events forward in time. This structural variety has generated substantive theoretical debate about the status of futurity in English. Current evidence indicates that future time reference in English emerges through interactions between modality and aspect rather than functioning as a separate grammatical category. Traditional approaches often position *will* as the standard or "neutral" (Leech, 1971, p.52) future marker, but research in cognitive linguistics and grammaticalization theory demonstrates that future reference operates through overlapping grammatical systems, where modal, aspectual, and temporal elements influence each other.

English future constructions have developed along specific historical paths, typically evolving from expressions of desire, obligation, or movement (Bybee et al., 1994). Hopper and Traugott's (2003) analysis of *going to* documents this evolution precisely, showing how a movement verb transformed into a future marker through identifiable stages of grammatical and semantic change. This pattern exemplifies broader processes in the development of temporal reference across languages. Building on Lyons' (1977) distinction between grammatical tense and semantic time, Palmer (1979, 1990) clarified how *will* functions within epistemic and deontic modal domains, revealing the specific relationships between modality and temporal reference in English.

Literature Review

The expression of future time in English emerges through interactions between multiple linguistic subsystems and learning mechanisms. While earlier research often examined isolated theoretical frameworks, recent usage-based grammar studies identify four key factors that shape future time expression in learner language: input frequency effects, processing constraints, L1 influence, and register-specific demands.

Input frequency significantly affects temporal marker acquisition. Corpus analyses of materials used with L2 learners (Collins, 2009; Mair, 2006) document that *will* constructions occur much more frequently than *be going to* in instructional materials and academic texts. This distribution directly shapes learner production patterns. Additionally, pedagogical materials typically present *will* as the primary future marker (Bardovi-Harlig, 2000), reinforcing these frequency-based acquisition patterns.

Processing constraints represent a second critical factor in temporal expression. The cognitive demands of real-time language production lead learners to prefer structurally simpler constructions. The *will + verb* structure constitutes a less complex syntactic pattern than the four-part *be going to* construction, reducing processing load during production (Tremblay et al., 2011). This processing advantage becomes particularly important in academic writing, where learners must allocate cognitive resources across content planning, argument structure, and register maintenance simultaneously.

L1 influence affects temporal expression through both facilitative and interfering effects. Languages differ substantially in how they mark future time—from dedicated morphological forms to lexical expression via temporal adverbials. These cross-linguistic differences shape how learners approach English future marking. Importantly, L1 influence rarely operates in isolation but interacts with other factors. For example, processing constraints often amplify L1 transfer when learners encounter structures that differ significantly from their native language patterns.

Register-specific demands constitute an often overlooked yet crucial factor in future time expression. Academic writing imposes specific constraints through its emphasis on formal style, explicit logical relationships, and precise temporal reference. These requirements affect both the frequency and distribution of future markers differently than in other discourse contexts. As a result, patterns observed in academic writing may not apply to other registers or modes of production.

Corpus-based research documents the prevalence of routinized, formulaic uses of such constructions. Schmitt (2004) notes that while learners readily acquire high-frequency chunks, they struggle with less transparent combinations, particularly tense-aspect patterns. Wood (2015) demonstrates how these patterns become conventionalized through repeated exposure, serving as "reliability islands" during early and intermediate development. Hoey's (2005) lexical priming research explains the mechanism behind this process, showing how repeated exposure to word combinations leads to their psychological entrenchment—a process especially evident in tense-aspect pattern acquisition.

The processing dimension of formulaic sequences provides additional insights. Ellis et al. (2008) demonstrate that chunked language enables faster and more accurate processing than novel combinations in both comprehension and production, with particular relevance for temporal expressions where conventional forms support fluent production. Eye-tracking studies by Conklin and Schmitt (2012) confirm these processing advantages in both native speakers and advanced learners, though with stronger effects in native speakers. Register variation adds further complexity to formulaic sequence acquisition. Biber and Barbieri's (2007) lexical bundle research maps how different registers favor specific formulaic sequence types, revealing specific patterns between register constraints and conventional expression distribution. Chen and Baker's (2010) analysis of academic writing identifies systematic differences between native and learner use of formulaic language, documenting distinct patterns of overuse and underuse. These findings explain why mastering register-appropriate formulaic language challenges even advanced learners.

Despite numerous studies analyzing formulaic sequences in learner language, English future tense remains understudied in second language acquisition research, despite its theoretical significance and position in the English tense system. Bardovi-Harlig's (2002, 2004) studies represent important exceptions, connecting theoretical aspects of English future tense with formulaic sequences by examining future expressions in learner language. These studies show that future expression emerges early in L2 development, first appearing only through *will*, with *be going to* developing later. Longitudinal data reveals that *be going to* initially appears in formulaic constructions as described by Ellis

(1997), with creative production developing subsequently. This developmental sequence persists regardless of instructional input, as learners continue to prefer *will* even when taught *be going to* earlier. Analysis of L2 production (Bardovi-Harlig, 2004) documents higher frequencies of *will* compared to *be going to* than in native speaker discourse. While learners use other future forms, *will* remains predominant, highlighting differences between native and non-native future time expression. These patterns stem from several factors: the modal properties of English future expressions, the range of context-dependent future forms available, and L2-specific processing constraints. L2 learners typically select simpler forms across future contexts, often overlooking the semantic and pragmatic constraints that guide native speaker usage. As Bardovi-Harlig (2002, p.198) notes, "although the use of formulaic language seems to play a limited role in the expression of future, its influence is noteworthy."

Research Aim and Research Questions

With a similar framework, this study examines how learners with different first languages use *will* and *be going to* in written English. Through corpus analysis, it explores both the frequency and patterns of these future time markers in learner writing. The analysis focuses on whether first languages influence how learners acquire and use English future expressions. Previous corpus studies have documented differences between native and non-native use of future expressions, yet the role of learners' first languages remains unexplored. By examining learner corpora across multiple first language backgrounds, this study addresses three primary research questions:

1. What is the distribution of *will* and *be going to* constructions in the learner corpus?
2. How does the use of these future time expressions vary across different L1 backgrounds?
3. What structural patterns, if any, emerge in the use of future time expressions within specific L1 groups?

To address these questions, the study employs a mixed-methods approach combining quantitative corpus analysis with qualitative examination of structural patterns. This methodology enables both broad distributional insights and detailed analysis of how learners from different L1 backgrounds deploy future time expressions in their writing.

Methodology

Data Collection

The International Corpus of Learner English (ICLE), version 3, compiled by Granger et al. (2020), constitutes a substantial resource for second language acquisition research, providing extensive data on non-native English language production. The corpus contains 5.7 million words from 25 national subcorpora of academic writing by advanced English language learners at the B2 and C1 proficiency levels of the Common European

Framework of Reference for Languages (CEFR). ICLE's methodological design enables fine-grained analyses of interlanguage patterns and developmental trajectories through the inclusion of metadata on learners' linguistic backgrounds, demographics, and educational contexts. ICLE's composition of academic essays written by university-level non-native English speakers allows for cross-linguistic comparisons across various national and linguistic contexts. The methodological strengths of ICLE lie in its systematic capture of learner language across a wide range of linguistic backgrounds and proficiency levels, as well as its extensive metadata.

Future-Tense Pattern Recognition and Extraction Framework

Future expressions were extracted using a theoretically-grounded approach based on contemporary analyses of English temporal construction patterns (Huddleston & Pullum, 2002; Biber et al., 1999). The process utilized the spaCy (Honnibal et al. 2020) natural language processing toolkit (version 3.7) with the *en_core_web_trf* model—a transformer-based architecture specifically optimized for analyzing complex syntactic dependencies.

The extraction protocol targeted two primary categories of future expressions. First, *will* expressions were identified through syntactic patterns centered on modal auxiliary usage. This process captured the complete constructional frame, including the subject (both nominal and pronominal forms), the modal auxiliary *will*, the main verb with its complements, and all associated modifiers and adjuncts. Specific processing for contracted forms (*'ll*) was implemented to ensure comprehensive coverage of all *will*-based future markers.

Second, *be going to* expressions were identified based on their semi-auxiliary constructional patterns. This extraction encompassed the subject element, the inflected form of *be* (*am*, *is*, *are*), the *going to* sequence, the main verb, and all associated complements and modifiers. This comprehensive approach captured the full range of structural variations in *be going to* constructions.

The extraction process was implemented through dependency parsing, utilizing spaCy's dependency labels and part-of-speech tags. Specifically, *will* expressions were identified through tokens carrying the modal verb (MD) tag and the *aux*dependency label. For *be going to* expressions, present participle forms (*going*) with appropriate auxiliary support and infinitival complements were located. This technical approach enabled precise identification of future expressions throughout the corpus.

Statistical Analysis

The analysis began by calculating the frequencies of verbs used in the extracted future time expressions, providing insight into the lexical preferences of L2 learners when constructing sentences with *will* and *be going to* structures. Beyond verb frequencies, the study examined the syntactic patterns of these expressions, including the distribution and combination of grammatical elements such as subjects, auxiliary verbs, and main verbs.

This examination aimed to identify tendencies or regularities in learners' use of future time expressions.

To measure the uniformity of future time expressions across the corpus, an entropy score was calculated for each lemmatized verb. Entropy quantifies how evenly a verb or expression pattern is distributed across different documents. The formula for entropy is based on the probabilities of a verb or pattern appearing in each document. Specifically, the entropy score is calculated as the negative sum of the product of the probability of the verb or pattern appearing in a document and the logarithm of that probability. A higher entropy value indicates that a verb or pattern is consistently employed across a wide range of documents, reflecting common lexical choices or syntactic patterns among learners. Conversely, a lower entropy score suggests that the verb or pattern's usage is concentrated in fewer documents, indicating infrequent application or context-specific usage.

The probability of each verb or pattern appearing in a document was calculated by dividing its frequency in that document by its total frequency across all documents. This ensures that the probabilities form a valid distribution for entropy calculation. Several factors were considered in the entropy calculation to ensure accuracy. Variations in corpus sizes, the number of documents (i.e., learners/writers), and the possibility of skewed distributions due to uneven instances of future time expressions across documents were accounted for. For example, a single document might contain a disproportionately high or low number of such expressions, which could distort results. To address this, the probability of each verb or pattern appearing across documents was determined by counting the number of unique documents in which each unit appears within each language subcorpus. These counts were then normalized by the total number of documents in each subcorpus, adjusting for differences in subcorpus sizes. This normalization ensures that the probabilities reflect the relative frequency of each unit within the context of each subcorpus, rather than absolute frequencies that could bias the analysis. This approach avoids relying solely on raw frequency counts, which may not accurately reflect L1 influence. Instead, it provides a more nuanced understanding of how verbs and expression patterns are distributed across learners and contexts, enabling a robust analysis of lexical and syntactic patterns in L2 future time expression usage.

Cluster analysis served as a complementary method to explore how different L1 groups pattern in their use of future time expressions. This analytical approach aims to identify groups of languages that share similarities in how speakers employ *will* and *be going to* constructions by examining formulaic patterns in their usage. The clustering methodology focuses on several features: the proportional use of each future expression type, subject patterns (pronouns, nouns, and proper nouns), and verb collocations. The analysis required several methodological controls to account for corpus size differences. Document counts were standardized to 200 per language (except Chinese with 127 documents due to data constraints). For each language, proportions of different patterns were first calculated within individual documents and then averaged across all documents in that language's subcorpus. These patterns include the relative frequencies of *will* versus *be going to*, the distribution of subject types, and the range of main verbs used with each future expression.

Effect sizes were calculated using Cohen's d to quantify the magnitude of differences between clusters. This measure was chosen over alternatives because it allows for meaningful comparison of differences across features with varying scales and distributions. The calculations accounted for different sample sizes between clusters using pooled standard deviation.

K-means clustering with k=4 was employed to identify statistically distinct groups based on these features. The silhouette score analysis helped evaluate the coherence of the resulting clusters, ensuring that the identified groups represent meaningful distinctions in how different L1 groups employ future time expressions. The choice of k=4 for clustering was supported by silhouette score analysis (0.72 for k=4 vs. 0.68 and 0.61 for k=2 and k=3 respectively) and stability measures across multiple runs. This methodological approach provides a quantitative means of examining how multiple features of future time expression usage combine and pattern across different L1 backgrounds. By considering multiple features simultaneously while maintaining statistical rigor through proper normalization and optimization techniques, the analysis examines broader patterns in how different L1 groups approach the expression of future time in English.

Publication Ethics

This research was undertaken in strict adherence to ethical guidelines governing research methodology and publication standards.

Results

The analysis of future time expressions across learner groups revealed systematic patterns of variation at multiple linguistic levels. The findings are presented in three complementary parts. The first part examines the distribution and entropy patterns of future time expressions across different L1 groups, focusing on the relative frequencies of *will* and *be going to* constructions. The second part analyzes the distinctive linguistic features that characterize different learner groups, including part-of-speech distributions and verb collocations. The final part explores the broader patterns of relationship between L1 backgrounds through principal component analysis and hierarchical clustering, revealing both macro-level distinctions and more subtle groupings among learner populations. The findings demonstrate substantial variation in how learners from different L1 backgrounds deploy future time expressions in English, with patterns emerging at both the broad typological level and within specific language families.

Table 1. Distribution and Entropy Patterns of Future Time Expressions across L1 Groups

Learner L1s	**Total Docs**	***will***		***be going to***		**Total**	**Mean per Doc**
		Freq	*Entropy*	*Freq*	*Entropy*	*Expr*	*Per Doc*
Chinese-Cantonese	746	3,051	9.214	86	6.310	3,137	4.21

Tswana	413	1,341	8.238	105	6.259	1,446	3.50
Swedish	388	1,337	8.213	54	5.030	1,391	3.59
Hungarian	315	1,046	7.891	10	2.846	1,056	3.35
Greek	312	678	7.887	41	4.634	719	2.30
German	312	816	7.874	13	3.700	829	2.66
Korean	289	765	7.850	23	4.230	788	2.73
Norwegian	280	943	7.758	41	5.016	984	3.51
Japanese	279	671	7.772	30	4.774	701	2.51
French	274	1,204	7.703	47	4.842	1,251	4.57
Russian	267	919	7.712	32	4.875	951	3.56
Persian	263	801	7.639	35	4.593	836	3.18
Portuguese	261	611	7.675	39	4.436	650	2.49
Polish	260	843	7.554	24	4.252	867	3.33
Serbian	259	771	7.647	37	4.551	808	3.12
Macedonian	248	945	7.496	33	4.203	978	3.94
Lithuanian	244	845	7.503	22	4.278	867	3.55
Italian	241	583	7.600	19	3.892	602	2.50
Bulgarian	230	703	7.509	13	2.815	716	3.11
Dutch	229	1,088	7.260	22	4.096	1,110	4.85
Finnish	213	686	7.325	35	4.605	721	3.38
Czech	206	727	7.254	21	4.202	748	3.63
Spanish	198	603	7.101	89	5.341	692	3.49
Turkish	198	575	7.184	26	4.162	601	3.04
Chinese	127	423	6.658	10	3.322	433	3.41

Table 1 presents the distribution of English future time expressions across learners from 25 different L1 backgrounds, comprising 7,160 texts. The analysis identified 23,237 future time expressions, with *will* constructions (n = 22,357; 96.2%) substantially outnumbering *be going to* constructions (n = 880; 3.8%) across all L1 groups.

A detailed examination of entropy scores reveals complex distributional patterns across both constructions. For *will*constructions, entropy scores ranged from 6.658 (Chinese L1) to 9.214 (Chinese-Cantonese L1), with 24 of 25 L1 groups exhibiting scores above 7.0. This high and consistent entropy pattern indicates that learners from most L1 backgrounds distribute *will* constructions broadly throughout their texts rather than concentrating them in specific sections. Particularly high entropy scores were observed among Chinese-Cantonese L1 (9.214), Tswana L1 (8.238), and Swedish L1 learners (8.213), suggesting especially uniform distribution of *will* in their writing.

In contrast, entropy scores for *be going to* displayed greater variability, ranging from 2.815 (Bulgarian L1) to 6.310 (Chinese-Cantonese L1). Five L1 groups - Chinese-Cantonese (6.310), Tswana (6.259), Spanish (5.341), Swedish (5.030), and Norwegian (5.016) - achieved entropy scores above 5.0. This pattern suggests more localized usage of *be going to*within texts, possibly indicating that learners employ this construction in specific contextual or rhetorical environments rather than throughout their writing. The markedly lower entropy scores among Hungarian L1 (2.846), Bulgarian L1 (2.815), and Chinese L1 (3.322) learners point to particularly concentrated usage patterns.

The frequency and density patterns complement these entropy findings. Chinese-Cantonese L1 learners, who showed the highest entropy scores for both constructions, also produced the highest frequency of *will* constructions (n = 3,051). Similarly, Tswana L1 learners demonstrated high entropy scores alongside the highest frequency of *be going*

*to*constructions (n = 105). The density of future expressions varied considerably, from Dutch L1 learners (4.85 expressions per text) to Greek L1 learners (2.30 expressions per text), suggesting that overall frequency of future marking may be independent of distribution patterns.

Table 2. Key Linguistic Features by Cluster for Future Expression Patterns

Feature Category and Type	Cluster 0 (n=24)	Cluster 1 (Chinese)	Mean Difference	Effect Size
Expression Patterns				
will	95.4	97.1	-1.6	0.84
be going to	4.6	2.9	1.6	0.76
Part-of-Speech Distribution				
Pronouns (PRON)	60.7	59.5	1.3	0.42
Proper Nouns (PROPN)	2.3	0.4	1.9	1.85
Common Nouns (NOUN)	36.9	40.1	-3.2	0.95
Combined Patterns				
will + PRON	60.4	58.7	1.8	0.55
will + PROPN	2.4	0.4	2.0	1.78
will + NOUN	37.2	41.0	-3.7	1.12
be going to + PRON	69.2	70.0	-0.8	0.38
be going to + PROPN	2.2	0.0	2.2	1.92
be going to + NOUN	28.5	30.0	-1.5	0.48
Top Distinctive Verbs				
have	4.6	2.0	2.6	1.24
be	8.7	13.7	-5.0	1.56
find	1.6	4.1	-2.6	1.32
discuss	0.7	2.2	-1.5	0.89
seem	0.8	2.2	-1.4	0.82
stay	0.2	1.5	-1.3	0.94
believe	1.6	2.8	-1.2	0.76
deny	0.1	1.2	-1.1	0.88
talk	0.2	1.3	-1.1	0.85
lose	0.4	1.5	-1.1	0.79

Legend: All values except Effect Size are percentages. Effect Size calculated using standardized mean difference. Cluster 0 includes 24 languages from diverse language families. Cluster 1 consists of Chinese only. PRON = pronouns; PROPN = proper nouns; NOUN = common nouns. Verbs are ordered by absolute magnitude of Mean Difference.

The cluster analysis revealed clear differences in how future time is expressed between Cluster 0 (comprising 24 languages from diverse families) and Cluster 1 (Chinese). Table 2 summarizes these differences across four main linguistic dimensions: basic expression patterns, part-of-speech distributions, combined patterns, and verb collocations. The statistical measures show strong evidence for these distinctions, with effect sizes ranging from moderate to large.

Regarding basic expression patterns, Chinese learners showed a slightly stronger preference for *will* constructions. They used *will* 97.1% of the time compared to 95.4% in Cluster 0. This corresponded with a lower use of *be going to*constructions (2.9% versus 4.6%). While these percentage differences might appear small, the effect sizes tell a

different story. The effect sizes of 0.84 for *will* and 0.76 for *be going to* indicate these differences are statistically meaningful and not due to chance.

When examining part-of-speech distributions, more substantial differences emerged, particularly in how learners used nouns. Cluster 0 used proper nouns (names of specific people, places, or organizations) in 2.3% of cases, while Chinese data showed only 0.4%—a difference that produced one of the largest effect sizes in the study (d = 1.85). In contrast, Chinese learners employed common nouns more frequently (40.1% compared to 36.9% in Cluster 0), reflected in another large effect size (d = 0.95). The use of pronouns, however, was relatively similar between the two groups (60.7% for Cluster 0 and 59.5% for Chinese), with a moderate effect size (d = 0.42).

The differences became even more pronounced when analyzing combined patterns—how future markers pair with different parts of speech. For *will* constructions, Cluster 0 combined proper nouns at a rate of 2.4%, while Chinese data showed only 0.4%, resulting in a large effect size (d = 1.78). Even more striking, in *be going to* constructions, there were no proper noun combinations whatsoever in the Chinese data, producing the largest effect size in the entire study (d = 1.92). The differences in pronoun combinations were less dramatic, with effect sizes of 0.55 for *will* and 0.38 for *be going to* constructions.

Verb collocations—which verbs frequently appear with future markers—showed additional substantial differences between the clusters. Chinese learners used the verb "be" much more frequently (13.7% versus 8.7% in Cluster 0) and "find" at a rate of 4.1% compared to 1.6% in Cluster 0. These differences were supported by strong effect sizes of 1.56 and 1.32 respectively. Conversely, Cluster 0 demonstrated a higher frequency of the verb "have" (4.6% versus 2.0%, d = 1.24). The pattern extended to several other verbs that Chinese learners used more frequently. For example, "discuss" appeared 2.2% of the time in Chinese data compared to just 0.7% in Cluster 0 (d = 0.89). Similarly, Chinese learners used "seem" more often (2.2% versus 0.8%, d = 0.82) and "stay" (1.5% versus 0.2%, d = 0.94). The same trend continued with "believe" (2.8% versus 1.6%, d = 0.76), "deny" (1.2% versus 0.1%, d = 0.88), "talk" (1.3% versus 0.2%, d = 0.85), and "lose" (1.5% versus 0.4%, d = 0.79). These detailed verb usage patterns further support the finding that learners from different language backgrounds employ distinct strategies when marking future time in their academic writing.

Figure 1. Distribution of Languages in Principal Component Space: Future Expression Analysis

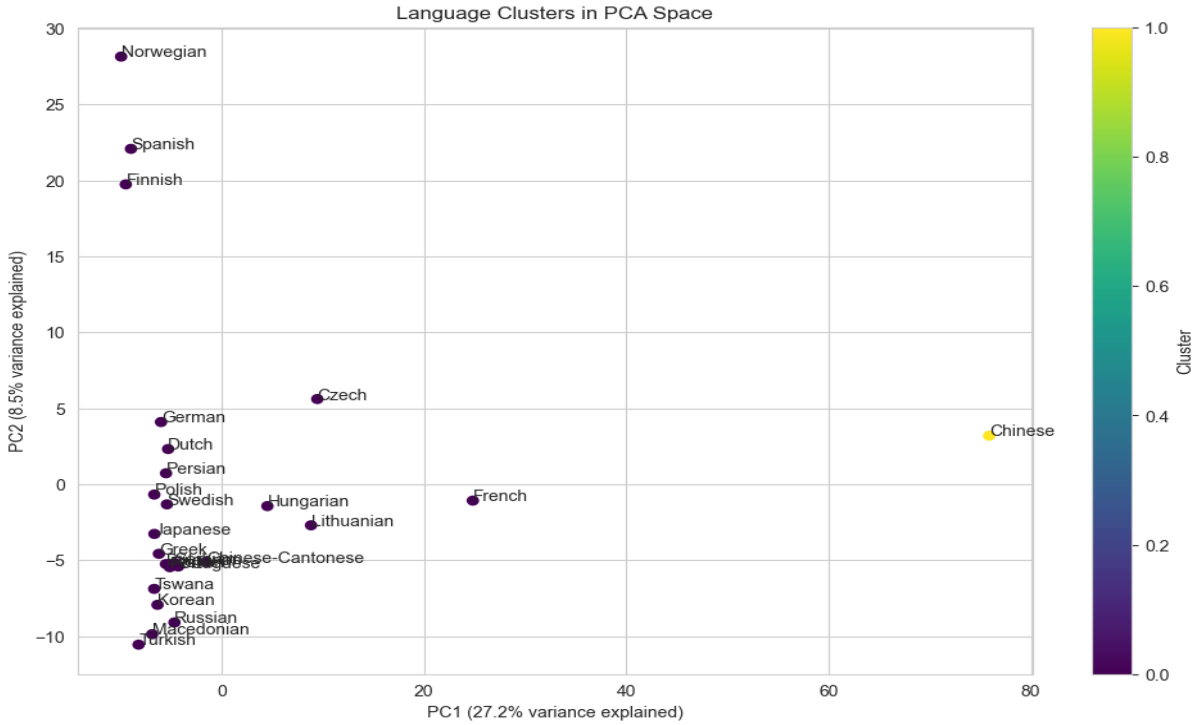
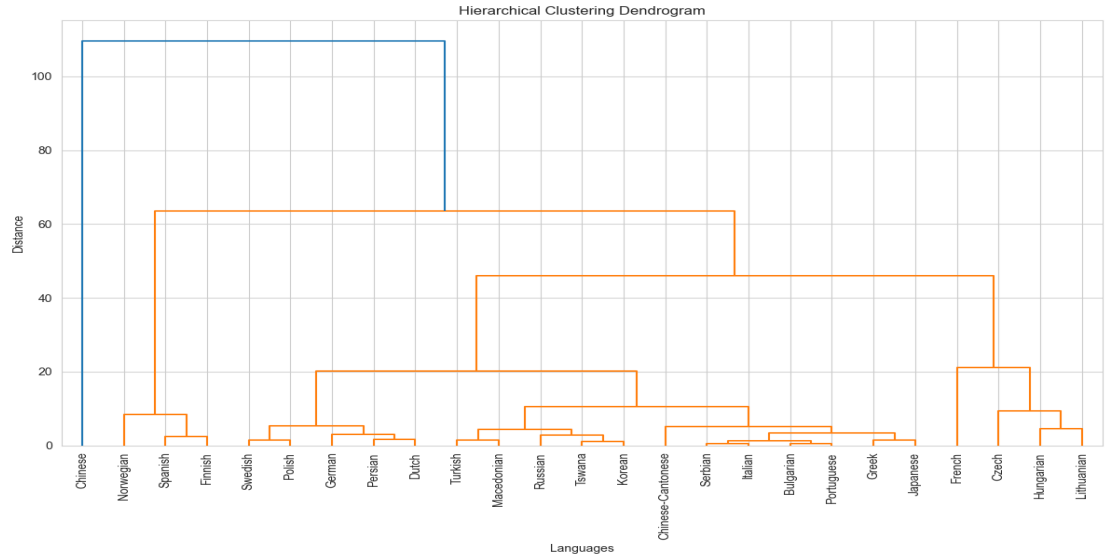


Figure 1 visualizes the Principal Component Analysis (PCA) of future expression patterns across 25 L1 backgrounds. The first two principal components explain 35.7% of the total variance (PC1: 27.2%, PC2: 8.5%). The most striking feature is the clear isolation of Chinese L1 learners (PC1 ≈ 80), suggesting fundamentally different patterns in their deployment of future time expressions.

The remaining 24 L1 groups cluster toward the negative end of PC1, with notable variation along PC2. The Nordic-Hispanic grouping (Norwegian, Finnish, and Spanish) forms a distinct cluster in the upper quadrant (PC2 > 15), while Slavic languages (Russian, Macedonian, and Bulgarian) demonstrate close grouping in the lower quadrant (PC2 < -5). Some languages occupy notable intermediate positions, with French showing relative isolation (PC2 ≈ 0, PC1 ≈ 20) and Czech displaying unique positioning relative to other Slavic languages.

The distribution pattern supports the two-cluster solution identified previously, with PC1 capturing the fundamental distinction between Chinese and non-Chinese L1 backgrounds, while PC2 reveals more nuanced variations among non-Chinese L1 groups.

Figure 2. Hierarchical Clustering of Future Expression Patterns Across Languages



The hierarchical clustering analysis (Figure 2) reveals nested relationships between L1 groups based on future expression patterns. The dendrogram's vertical axis represents the distance measure between clusters, with higher values indicating greater dissimilarity. Chinese branches early at approximately distance 100, confirming its distinct status.

Below the 60-distance threshold, the remaining L1 groups form coherent subgroupings. The Nordic-Hispanic cluster emerges clearly, with Norwegian and Spanish forming a tight cluster later joined by Finnish. Other notable groupings include the Czech-French pairing and a distinct Turkish-Dutch cluster. Slavic L1 backgrounds (Russian, Macedonian, Bulgarian) demonstrate tight clustering at relatively low distances, indicating shared patterns in future expression.

The complementary PCA and hierarchical clustering analyses provide converging evidence for systematic variation in future marking. While the PCA captures the primary Chinese/non-Chinese distinction (PC1: 27.2% variance), the dendrogram reveals finer relationships among L1 groups. These structural relationships align with linguistic features from Table 2, where Chinese exhibits distinctive patterns in verb usage ('be': 13.7% versus 8.7%; 'find': 4.1% versus 1.6%) and nominal constructions (proper nouns: 0.4% versus 2.3%).

This multi-method analysis reveals a layered structure in future expression patterns, where L1 groups maintain distinct characteristics partially aligned with genetic relationships between languages. These patterns manifest consistently across multiple linguistic features, suggesting systematic variation in how learners conceptualize and express future time in English.

Discussion

The analysis of future time expressions in learner English reveals several significant patterns in temporal reference acquisition. The overwhelming preference for *will* constructions (96.2%) over *be going to* (3.8%) across all L1 groups represents a striking finding that both confirms and challenges existing research. This marked imbalance in future marking, consistent across typologically diverse L1 backgrounds, supports Bardovi-Harlig's (2002) observation about limited temporal patterns in learner language. However, the cross-sectional data indicates that this strong preference for *will* persists even at advanced proficiency levels, extending Collins' (2009) findings on temporal marking stability.

The emergence of distinct cross-linguistic patterns, particularly among Chinese learners as shown in the principal component analysis (PC1=27.2% variance explained), provides new evidence for L1 influence while suggesting more complex interactions between language systems. The patterns in Chinese learners' production, particularly their higher frequencies of basic verbs like 'be' (13.7% versus 8.7%) and 'find' (4.1% versus 1.6%), support Slobin's (1996) concept of "thinking for speaking"—the idea that language shapes how speakers conceptualize events for expression—while indicating fundamental differences in how temporal concepts are encoded across languages.

The distribution patterns of future expressions, revealed through entropy analysis, demonstrate systematic differences beyond simple frequency effects. The consistently high entropy scores for *will* constructions (>7.0 in 24 of 25 languages) contrast with the restricted distribution of *be going to* (five languages showing entropy scores above 5.0), suggesting specialized pragmatic functions in learner discourse as described by Hopper and Traugott (2003). The hierarchical cluster analysis reveals systematic language groupings, particularly among Nordic and Slavic language families. These groupings support Ringbom's (2007) concept of "perceived linguistic distance"—how learners intuitively assess similarities between languages—while indicating that genetic relationships between languages may shape temporal conceptualization in previously unconsidered ways.

Future Expression Choice and Processing Constraints

The consistent preference for *will* constructions across L1 groups reflects fundamental constraints in second language processing. Skehan and Foster (2001) argue that second language learners must carefully manage their cognitive resources during production—similar to how computers allocate memory and processing power. The findings support this view, showing that learners favor simpler grammatical forms like *will* + verb over the more complex *be going to* when expressing future time. This observation aligns with VanPatten's (2002) research demonstrating learners' tendency to choose forms requiring minimal mental effort during communication.

The entropy analysis reveals distinct patterns in how future expressions are distributed throughout texts. The high entropy scores for *will* constructions (M=7.673) indicate learners' ability to access and use these forms readily throughout their writing,

which DeKeyser (2015) associates with easier processing through repeated use. In practical terms, this means learners can deploy *will* constructions fluently in various contexts within their texts. Conversely, the lower entropy scores for *be going to* constructions (M=4.432) reflect more selective usage of this complex form, appearing in more limited contexts. This pattern supports Robinson's (2005) findings on how grammatically complex structures demand more attention resources from learners.

The patterns in verb choice further illuminate processing strategies in learner writing. Chinese learners demonstrate a marked preference for basic verbs like 'be' and 'find'—verbs that serve as fundamental building blocks in language. François and Albakry (2021) suggest this simplification strategy operates systematically across L1 groups, though with varying intensity depending on language background. As Bardovi-Harlig (2004) noted, these patterns likely reflect both processing limitations and first language influence, particularly when grammatical structures differ significantly from learners' first languages. This interaction between cognitive constraints and language transfer creates distinctive patterns in how learners from different backgrounds approach future time expression.

Cross-linguistic Influence Patterns

The analysis reveals that a learner's first language shapes their expression of future time in English in several distinct ways. The principal component analysis shows a clear separation of Chinese learners from other groups, accounting for 27.2% of the variance in the data—meaning over a quarter of all differences observed can be attributed to this language distinction alone. Supporting Salaberry and Comajoan's (2002) argument that learners from different language backgrounds conceptualize time differently, Chinese learners demonstrate distinct patterns in verb usage ('be': 13.7% versus 8.7%; 'find': 4.1% versus 1.6%) and noun preferences (proper nouns: 0.4% versus 2.3%; common nouns: 40.1% versus 36.9%). These differences likely stem from how Chinese grammaticalizes temporal concepts in ways fundamentally different from Indo-European languages.

The hierarchical clustering analysis reveals systematic groupings of related languages, with Nordic languages (Norwegian, Finnish) and Slavic languages (Russian, Macedonian, Bulgarian) forming distinct clusters—essentially "family groups" based on future expression patterns. Extending Bardovi-Harlig's (2000) findings, these shared strategies persist even at advanced proficiency levels. As Ellis (2006) notes, L1 influence on temporal expression often remains visible after years of second language use, showing remarkable persistence compared to other language features.

The patterns extend to how learners combine future markers with different types of nouns. Chinese learners rarely combine future markers with proper nouns (*will* + proper nouns: 0.4% versus 2.4%; *be going to* + proper nouns: 0% versus 2.2%). Peters (2016) demonstrates that such differences often reflect language-specific strategies that learners transfer from their native language systems.

The entropy analysis reveals varying levels of consistency in future expression use across language backgrounds, particularly with *be going to* constructions (entropy scores

ranging from 2.815 to 6.310). As Jarvis (2000) shows, L1 influence appears in subtle ways, including how learners organize information within texts. These systematic differences indicate that L1 influences both the choice of future expressions and how these expressions are distributed throughout learners' writing.

The clustering patterns suggest, as Ringbom (2007) argues, that learners from related language backgrounds may share advantages or face similar challenges with English temporal expressions, rather than exhibiting simple one-to-one L1 transfer. Rather than direct transfer of specific structures, learners seem to develop approaches to future time expression that reflect broader typological similarities between their L1 and other languages. These effects persist even at advanced proficiency levels, suggesting they become entrenched aspects of learners' interlanguage systems.

Structural and Discourse Patterns

The distribution of future expressions reveals patterns at both structural and discourse levels, with frequency varying considerably across language groups (2.30 to 4.85 expressions per text)—more than a twofold difference in how frequently learners mark future time. Hopper and Traugott (2003) suggest such variation reflects different strategies for managing temporal reference in discourse, essentially different approaches to signaling when events will occur. As Klein (2009) notes, learners must balance temporal marking with other aspects of discourse organization, and the data shows they accomplish this in distinctly different ways depending on their language background.

The entropy analysis demonstrates consistent patterns in academic writing across language groups. The high entropy scores for *will* constructions (above 7.0 in 24 languages) indicate stable usage patterns—meaning learners use this form consistently throughout their texts rather than clustering it in certain sections. This finding supports Biber et al.'s (1999) research on how discourse patterns develop systematically in academic writing. In contrast, *be going to* shows a more restricted distribution (five languages with entropy scores above 5.0), aligning with Myhill's (1992) observation that different future markers serve distinct discourse functions, with some forms appearing only in specific contexts or for particular communicative purposes.

Verb choice patterns with future expressions provide additional insight into learner strategies. Chinese learners demonstrate strong preferences for certain verb combinations, building on Tagliamonte's (2006) research on systematic patterns in future expression. As Poplack and Tagliamonte (2000) note, such preferences often reflect learned strategies for managing temporal reference in specific contexts—essentially, learners develop standard ways to talk about the future that become habitual in their writing. Clancy (2016) found that learners develop context-specific time-marking approaches, evident in the analysis of academic writing presented here.

The distribution patterns suggest a gradual development of temporal reference systems, as described by Hilpert (2008) and Langacker (2008). Following Declerck's (2006) work on how temporal reference develops in language and Huddleston and Pullum's (2002) analysis of the relationship between modality and time expression, these

findings demonstrate how learners integrate these meanings in academic discourse. These patterns indicate that learners develop systematic strategies reflecting both their language background and understanding of academic conventions while managing the cognitive demands of second language production. Rather than random variation, the patterns reveal systematic approaches to future time expression that become established in learners' academic writing practices.

Formulaic Patterns in Future Expression

The analysis reveals strong evidence for the formulaic nature of future expressions in learner language. The strong preference for *will* (96.2%) across all groups suggests that learners develop what Wray (2002, 2008) calls prefabricated patterns—ready-made language chunks that can be retrieved and used as whole units—for efficient future time expression. The uniform distribution of *will* constructions (entropy scores above 7.0 in 24 languages) supports Wood's (2015) concept of "reliability islands" in learner language—stable phrases that learners can depend on when navigating the complexities of a second language.

Verb combinations with future markers further demonstrate formulaic language development. Supporting Hoey's (2005) research on lexical priming—the tendency for words to become associated with particular usage patterns through repeated exposure—learners from different language backgrounds show systematic preferences in verb choices. Ellis et al. (2008) demonstrate how learners develop stable form-meaning associations in formulaic sequences, where certain forms become strongly linked with specific meanings through repeated use. This phenomenon is evidenced in this study through consistent patterns in future expressions. Chinese learners' verb preferences ('be': 13.7%; 'find': 4.1%) align with Chen and Baker's (2010) findings on academic writing formulaic patterns and Sinclair's (1991) observations about how certain word combinations become conventionalized in language use.

The interaction between formulaic patterns and first language influence reveals distinct patterns in noun usage with future expressions (proper nouns: 2.3% versus 0.4% for Chinese learners; common nouns: 36.9% versus 40.1%). Pawley and Syder (1983) suggest L1 influence on formulaic sequence selection, where native language patterns affect which formulaic expressions learners adopt. Meanwhile, Kecskes (2007) describes formulaic language development as a dual-language process—where both first and second language systems interact during learning—creating systematic variations in how formulaic language is acquired and used.

The hierarchical clustering results demonstrate consistent structural patterns across language backgrounds, supporting Biber and Barbieri's (2007) research on standardized formulaic sequences in academic writing. The restricted use of *be going to* (five languages with entropy scores above 5.0) aligns with Wulff's (2018) findings on context-specific usage of less common formulaic sequences, suggesting that some patterns develop only for specific communicative contexts.

These findings indicate that while learners develop similar formulaic patterns for future time expression, their first languages influence pattern implementation, demonstrating a learning process that balances universal cognitive constraints with L1 influence. Rather than learning each grammatical rule separately, learners appear to acquire and deploy future time expressions as pre-packaged units, but with variations shaped by their native language backgrounds.

Theoretical Implications

This study advances the understanding of how learners acquire and use future time expressions in English. Previous research has emphasized first language influence in tense-aspect acquisition (Bardovi-Harlig, 2000; Salaberry & Comajoan, 2002)—the idea that how learners express time in English depends heavily on their native language patterns. However, the findings reveal patterns that cross language boundaries, suggesting that universal factors—cognitive processes common to all language learners—play a larger role than previously thought. The strong preference for certain future expressions across diverse language backgrounds points to common developmental pathways in how learners come to mark future time in English.

The results shed new light on how universal patterns interact with first language influence. While the dominance of *will*constructions (96.2% across all language groups) supports Wray's (2008) view that learners develop formulaic language based on communicative needs, the variations in verb choices and noun usage suggest more subtle effects of first language influence. These patterns show that first language influence shapes how learners implement common strategies rather than determining their basic approach to future marking. Rather than a simple case of transfer or universal development, learners appear to follow shared cognitive pathways while expressing language-specific variations in implementation.

The entropy analysis offers a new way to understand formulaic language development—how learners acquire and use conventional phrases and patterns. Ellis et al. (2008) proposed that learners acquire formulaic sequences in a relatively straightforward progression from fixed chunks to more creative usage. However, the analysis of how *will* and *be going to*are distributed in texts suggests a more complex developmental pattern. The consistently high entropy scores for *will*(above 7.0 in 24 languages) compared to the more variable and generally lower scores for *be going to* (five languages above 5.0) indicate different acquisition and usage patterns for these constructions. Learners appear to develop systematic ways of using future expressions that reflect both universal tendencies in language processing and specific influences from their first language, creating a more nuanced picture of formulaic language development than previously understood.

Pedagogical Implications

The findings suggest several ways to improve the teaching of future time expressions in English language classrooms. While current textbooks often treat *will* and *be going to* as equally important alternatives (Biber et al., 1999), the research demonstrates learners' natural preference for *will* constructions. Teaching materials could

acknowledge this preference while gradually introducing other future expressions. For example, instructors might first focus on solidifying *will* usage in various contexts before introducing the more complex *be going to* construction, aligning pedagogy with natural acquisition patterns.

The distinct patterns among language groups, particularly Chinese learners, indicate the value of linguistically-informed teaching approaches. These could include targeted exercises addressing specific challenges, such as expanding verb range with future expressions and practicing varied noun usage in future constructions. For instance, Chinese-speaking learners might benefit from focused practice combining future markers with proper nouns, an area where the data shows particular divergence from other language groups (0.4% versus 2.4% for *will* with proper nouns).

The entropy analysis reveals implications for academic writing instruction. The systematic differences in future expression distribution suggest focusing on how temporal markers contribute to text organization, moving beyond sentence-level grammar to teach students how to create coherence throughout their writing. Rather than treating future expressions as interchangeable grammatical forms, instructors could help students understand how different future markers serve distinct discourse functions and contribute to the overall structure and flow of academic texts. This approach would connect grammar instruction to broader skills in academic writing organization and reader expectations in English academic discourse.

Conclusion

This study has examined how learners from 25 different language backgrounds express future time in English academic writing. Through frequency analysis, entropy measures that quantify distribution patterns, and multivariate statistical techniques that identify relationships across multiple variables, the research provides new insights into the acquisition and use of English future expressions.

Three key findings emerged from this comprehensive analysis: First, learners across all language backgrounds strongly prefer *will* constructions (96.2%), suggesting common developmental patterns in how future time marking evolves in second language acquisition. Second, the entropy analysis revealed different distribution patterns, with *will* appearing more uniformly throughout texts than *be going to*, which tends to occur in more restricted contexts. Third, the multivariate analyses identified distinct patterns among language groups, with Chinese learners showing particularly distinctive strategies that set them apart from other language backgrounds.

These findings advance our understanding of temporal expression development in second language acquisition, demonstrating the complex interaction between universal cognitive patterns and first language influence. The results suggest practical teaching approaches that recognize both common developmental trajectories and group-specific challenges faced by learners from particular language backgrounds. Future research directions include longitudinal studies tracking how these patterns develop over time in individual learners, investigations of how different registers (such as conversation versus academic writing) affect future expression choices, and examination of the relationship

between how learners comprehend future expressions and how they produce them in their own writing and speaking.

References

Bardovi-Harlig, K. (2000). *Tense and aspect in second language acquisition: Form, meaning, and use*. Blackwell.

Bardovi-Harlig, K. (2002). A new starting point? Investigating formulaic use and input in future expression. *Studies in Second Language Acquisition, 24*(2), 189–198.

Bardovi-Harlig, K. (2004). The future in the past: The acquisition of the present progressive with future time reference. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133–154). John Benjamins.

Berglund, Y. (1997). Future in present-day English: Corpus-based evidence on the rivalry of will and be going to. In M. Ljung (Ed.), *Corpus-based studies in English: Papers from the seventeenth international conference on English language research on computerized corpora (ICAME 17)* (pp. 31–43). Rodopi.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263–286.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.

Bishop, J. (2018). Processing constraints on L2 formulaic sequence acquisition: Evidence from a corpus of learner speech. *Language Learning, 68*(2), 465–495.

Bybee, J., Perkins, R., & Pagliuca, W. (1994). *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. University of Chicago Press.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30–49.

Clancy, B. (2016). Prosody and the disambiguation of future temporal reference in English conversation. *Journal of Pragmatics, 95*, 1–16.

Collins, P. (2009). *Modals and quasi-modals in English*. Rodopi.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics, 32*, 45–61.

Declerck, R. (2006). *The grammar of the English tense system: A comprehensive analysis*. Mouton de Gruyter.

DeKeyser, R. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 97–114). Routledge.

Ellis, N. C. (1997). Vocabulary acquisition: Word structure, collocation, and meaning. In M. McCarthy & N. Schmidt (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 122–139). Cambridge University Press.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2), 143–188.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics, 27*(2), 164–194.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*(3), 375–396.

François, T., & Albakry, M. (2021). Strategic simplification in L2 production: Evidence from temporal reference. *Studies in Second Language Acquisition, 43*(1), 1–25.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2020). *The international corpus of learner English* (Version 3). Presses universitaires de Louvain.

Hilpert, M. (2008). The future of "going to": A corpus-based diachronic analysis. In A. Ziegler (Ed.), *Corpora, creativity, and cognition: Theory, method, and practice* (pp. 133–154). Narr.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength natural language processing in Python*. https://spacy.io/

Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization* (2nd ed.). Cambridge University Press.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge University Press.

Jarvis, S. (2000). Methodological rigor in the study of transfer: Identifying L1 influence in the interlanguage lexicon. *Language Learning, 50*(2), 245–309.

Kecskes, I. (2007). Formulaic language in English lingua franca. In I. Kecskes & L. Horn (Eds.), *Explorations in pragmatics: Linguistic, cognitive and intercultural aspects* (pp. 191–219). Mouton de Gruyter.

Kellerman, E. (1995). Crosslinguistic influence: Transfer to nowhere? *Annual Review of Applied Linguistics, 15*, 125–150.

Klein, W. (2009). How time is encoded. In W. Klein & P. Li (Eds.), *The expression of time* (pp. 39–82). Mouton de Gruyter.

Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. Oxford University Press.

Leech, G. (1971). *Meaning and the English verb*. Longman.

Lyons, J. (1977). *Semantics* (Vols. 1–2). Cambridge University Press.

Mair, C. (2006). *Twentieth-century English: History, variation and standardization*. Cambridge University Press.

Myhill, J. (1992). *Typological discourse analysis: The language of evaluation*. Continuum.

Odlin, T. (2003). Cross-linguistic influence. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 436–486). Blackwell.

Palmer, F. R. (1979). *Modality and the English modals*. Longman.

Palmer, F. R. (1990). *Mood and modality*. Cambridge University Press.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–226). Longman.

Peters, E. (2016). Formulaic language in L2 acquisition: The effects of frequency, semantic transparency, and type of target language. *Language Learning, 66*(2), 303–333.

Poplack, S., & Tagliamonte, S. (2000). The grammaticization of going to in (African American) English. *Language Variation and Change, 11*(3), 315–342.

Ringbom, H. (2007). *Cross-linguistic similarity in foreign language learning*. Multilingual Matters.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *IRAL – International Review of Applied Linguistics in Language Teaching, 43*(1), 1–32.

Salaberry, R., & Comajoan, L. (2002). The acquisition of English tense-aspect morphology: A generative perspective. In R. Salaberry & Y. Shirai (Eds.), *The L2 acquisition of tense-aspect morphology* (pp. 1–31). John Benjamins.

Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing, and use*. John Benjamins.

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

Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction*(pp. 183–205). Cambridge University Press.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking." In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). Cambridge University Press.

Tagliamonte, S. (2006). *Analysing sociolinguistic variation*. Cambridge University Press.

Tremblay, A., Derwing, B., & Libben, G. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall. *Language Learning, 61*(2), 569–613.

VanPatten, B. (2002). Processing instruction: An update. *Language Learning, 52*(4), 755–803.

Wood, D. (2015). *Fundamentals of formulaic language: An introduction*. Bloomsbury.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford University Press.

Wulff, S. (2018). Formulaicity in L2 acquisition. In P. Booth & J. Swann (Eds.), *The Routledge handbook of second language acquisition* (pp. 442–456). Routledge.