Compositional Syntactico-SemBanking for English as a Second or Foreign Language

Wenxi Li[1,2], Xihao Wang[3], Weiwei Sun[4]

[1]School of the Chinese Nation Studies, Minzu University of China
[2]Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China
[3]Department of Chinese Language and Literature, Peking University
[4]Department of Computer Science and Technology, University of Cambridge

Correspondence: ws390@cam.ac.uk

Abstract

Despite the widespread use of English as a Second or Foreign Language (ESFL), developing syntactico-semantic representations for it is limited — the irregularities in ESFL complicate systematic composition and subsequently the derivation of its semantics. This paper draws on constructivism and proposes a novel Synchronous Hyperedge Replacement Grammar (SHRG)-based constructivist approach to address the challenges. By using constructions as fundamental units, this approach not only accommodates both the idiosyncrasies and the compositional nature of ESFL, but also bridges the gap between literal cues and intended meaning. The feasibility of this constructivist approach is demonstrated using real ESFL data, resulting in a gold-standard, medium-sized syntactico-semantic bank that covers a wide range of ESFL phenomena.

1 Introduction

The human language faculty is widely believed to consist of two key components: (i) a *lexicon* of stored, memorized listmes and (ii) a *mental grammar* that defines how novel utterances are composed from them (Jackendoff, 2000, 2002). Guided by the principle of compositionality, which holds that the meaning of an expression arises from the meanings of its parts and their syntactic combination (Partee, 1984), these components are thought to enable humans to systematically generate or comprehend an infinite number of expressions.

However, such a compositional approach faces significant challenges in characterizing the syntax–semantics interface of non-native languages. The frequent presence of irregularities in them, including unconventional tokens and structures, makes it exceedingly difficult to define their lexicons and grammars. Yet, speakers consistently demonstrate the ability to interpret these languages effectively (Clahsen and Felser, 2006; Kurz, 2008). Consider sentences 1a–1c which are real-world examples of non-native English, i.e., English as a Second or Foreign Language (ESFL)[1]. Despite containing apparent syntactic errors, each sentence successfully conveys its intended semantics:

(1) a. Omission: *I had to sleep in **(a)** tent.*

b. Insertion: *We contacted ~~**with**~~ Kim.*

c. Transposition: *He **visits often** Paris.*

How can coherent semantics arise from an "imperfect" syntactic foundation? This question has become increasingly pressing, as ESFL, with the advent of globalization, has come to be recognized not merely as an erroneous forms of native English (Bryant et al., 2023), but as a linguistic system in its own right (Khanuja et al., 2020; Aguilar and Solorio, 2020; Zhang et al., 2021; Bryant et al., 2023). Notably, over 70% of English speakers worldwide are non-native (Eberhard et al., 2024).

In response, this paper draws on insights from constructivist linguistics to address this challenge. Specifically, grounded in the constructivist theories which conceptualize words, idioms, and phrases as form–meaning pairings, or constructions, stored in an inventory that exhaustively captures human language (Bates and MacWhinney, 1987; Langacker, 1987; Goldberg, 1995, 2003; Croft, 2001; Croft and Cruse, 2004; Tomasello, 2003; Robinson and Ellis, 2008), this paper proposes using syntactico-semantic constructions as the fundamental units for modeling the computational systems underlying the human language faculty.

As illustrated in Figure 1, this approach treats diverse "errors" in ESFL as distinct form–meaning pairings. In doing so, it not only accommodates these deviations but also integrates them into constructional templates, thereby enabling their participation in broader syntactico-semantic composition.

[1]Data are from the Cambridge First Certificate in English dataset (FCE; Yannakoudakis et al., 2011)

Findings of the Association for Computational Linguistics: ACL 2025, pages 24395–24406

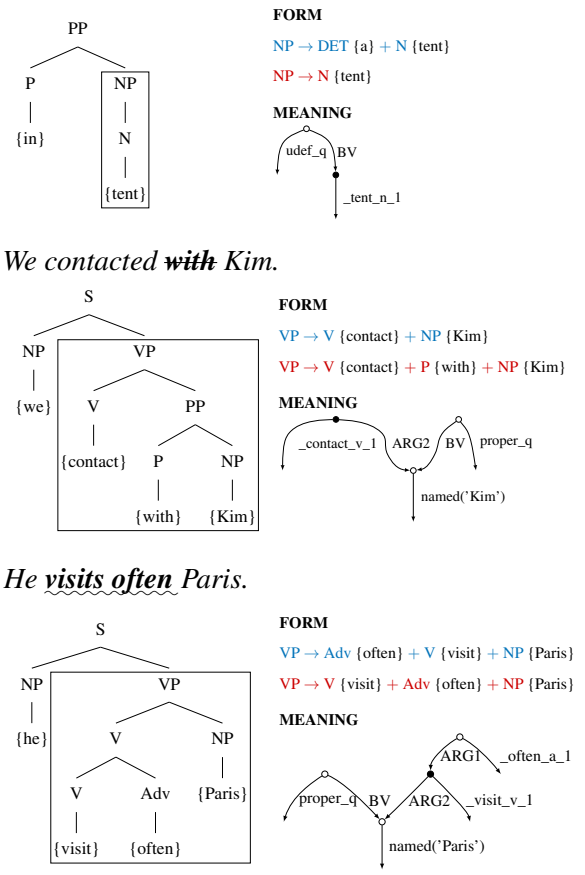


Figure 1: Constructivist analysis of Example 1a – 1c. The left side shows where the "errors" are embedded and the right side presents how constructions — FORMs (blue for corrected versions, red for ESFL data) paired with MEANINGs — accommodate these "errors".

At the implementation level, we represent the syntax of ESFL using Context Free Grammar (CFG) trees and the semantics using graphs in the framework of English Resource Semantics (ERS; Flickinger et al., 2014a). To model the parallel composition of syntactic and semantic representations, Synchronous Hyperedge Replacement Grammar (SHRG), a synchronous extension of Hypergraph Replacement Grammar (HRG) that has been widely applied in semantics-based parsing and generation for English (Chen et al., 2018; Chen and Sun, 2020; Ye and Sun, 2020), is adopted. Details of the constructivist syntactico–semantic formalism is presented in §2.

The feasibility and the effectiveness of the SHRG-based constructivist syntax–semantics interface is tested on real ESFL data. Following the workflow demonstrated in Figure 2, we develop a syntactico-semantic bank (sembank; §3), which achieves high inter-annotator agreement, addresses various phenomena effectively, and extends coverage to data previously deemed unparsable.

We argue that our newly-developed sembank for ESFL benefits both linguistic theory and practical applications. Theoretically, it shows that ESFL is compositional — not a random collection of expressions, but a natural language governed by systematic rules (Adjemian, 1976). Its derivational process parallels that of native English, accounting for its generativity. Practically, since language acquisition involves mastering form–meaning pairings of the target language, our sembank serves as a valuable resource for empirical research in this domain (Granger, 2003, 2014; Biber and Reppen, 2015), offering a deeper and more comprehensive analysis than existing ESFL corpora (MacWhinney, 2000; Sagae et al., 2007; Geertzen et al., 2013).

2 A Constructivist Approach

2.1 Theoretical Background

Though being a field encompassing a wide range of approaches (see Hoffmann and Trousdale, 2013; Goldberg, 2013), constructivist theories share a belief that human languages could be exhaustively described by interconnected constructions which integrate forms and meanings and are stored in an inventory (Bates and MacWhinney, 1987; Langacker, 1987; Goldberg, 1995, 2003; Croft, 2001; Croft and Cruse, 2004; Tomasello, 2003; Robinson and Ellis, 2008). These constructions, which relate the observable properties of their morphological, lexical, or syntactic forms to specific semantic, pragmatic, or discourse functions, exist at multiple levels of linguistic descriptions — morphemes, words, idioms, as well as phrasal constructions. In this way, the distinction between lexicon and grammar is blurred — both of them are viewed as learnable symbolic links between forms and meanings, emerging and evolving in specific contexts.

We argue that these features of constructivism make it particularly well-suited to addressing the challenges of modeling syntactico-semantic derivation in ESFL. First, the constructivist approach offers a framework for understanding how ESFL speakers compose meaning. By addressing irregularities through a continually expanding inventory of form–meaning pairings and using constructions — spanning varying levels of complexity and generality — as structural templates, it accommodates both the idiosyncrasies and the compositionality of

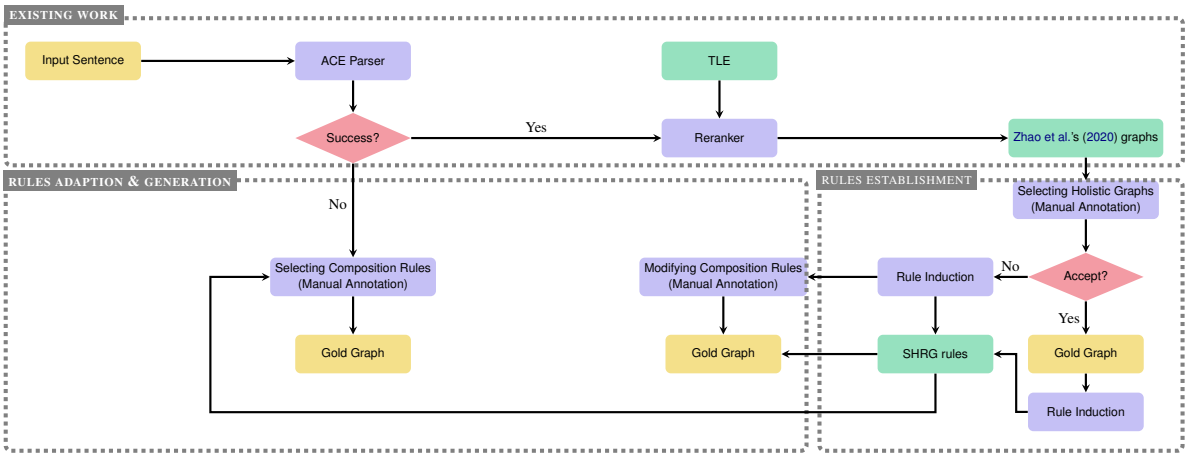


Figure 2: Construction of ESFL sembank, including input/output data (in yellow), associated sources (in green), technical operations (in blue) and workflow.

ESFL. Additionally, as a context-sensitive framework, the constructivism encourages a more flexible capturing of ESFL semantics, allowing it to emerge naturally from real usage. This fosters greater tolerance for so-called "errors" in ESFL.

2.2 Formalism

The Synchronous Hyperedge Replacement Grammar (SHRG), which captures mappings between two structured representations, is employed to model the constructivist syntax–semantics interface. Each construction is represented by an SHRG rule, which consists of a Context Free Grammar (CFG) for syntax and an Hyperedge Replacement Grammar (HRG) for semantics. The rewriting formalism of SHRG models how syntax and semantics are constructed synchronously.

Context Free Grammar CFG is utilized to represent syntactic constructions. Formally speaking, it is a 4-tuple $C = (N, \Sigma, P, S)$: (i) N is a finite set of non-terminal symbols; (ii) Σ is a finite set of terminal symbols that are disjoint from N; (iii) P is a finite set of rewriting rules in the form $A \rightarrow \beta$, where $A \in N$ and β is a string of terminals or non-terminals (β can be empty); (iv) S is the start symbol and $S \in N$. The derivation of the syntax can be done by successively applying the rewriting rules in the set P to non-terminal nodes.

Edge-labeled and Directed Hypergraph Semantics is represented by the edge-labeled and directed hypergraph. It is formally defined as $G = (V, E, L, l, X)$, where (i) V is a finite set of vertices; (ii) E is a finite set of hyperedges and each hyperedge is represented as an ordered tuple of vertices from V; (iii) L is a finite set of labels; (iv) l is a mapping from E to the finite set L, denoted as $l : E \rightarrow L$; (v) $X \in V^*$ defines an ordered list of nodes called external nodes, which specify the docking points during graph rewriting.

Hyperedge Replacement Grammar HRG specifies the stepwise substitution of non-terminal symbols in the hypergraph with corresponding hypergraphs. Formally speaking, HRG is a 4-tuple $H = (N, \Sigma, P, S)$: (i) N is a finite set of non-terminal symbols; (ii) Σ is a finite set of terminal symbols (disjoint from N); (iii) P is a finite set of hyperedge replacement rules, each rule of the form $A \rightarrow G$, where $A \in N$, and G is a hypergraph with edge labels over $N \cup \Sigma$ (G can be empty); (iv) S is the start symbol and $S \in N$.

Synchronous Hyperedge Replacement Grammar HRG can be extended to SHRG which provides a formal mechanism governing both the individual replacement of each structured representation and their synchronization. SHRG is defined as a 6-tuple $R = (N, \Sigma, P, S, Q, \Delta)$, and compared to HRG, its new elements include (i) Q is a finite set of non-terminal symbols for the second structured representation and (ii) Δ is a set of synchronization rules, specifying the replacement rules in the two structures and how these rules are related.

2.3 Informal Illustration

We illustrate our SHRG-based syntax—semantics interface with Example 1c. Figure 3 presents its form, the CFG tree C, alongside its meaning, represented as an edge-labeled directed hypergraph H following the ERS framework.

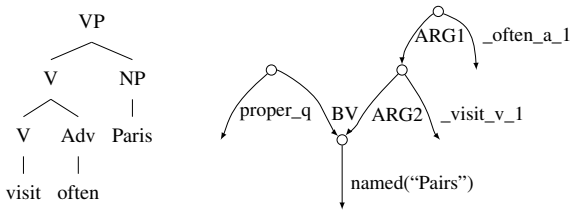


Figure 3: Form and meaning of *visit often Paris*

Table 1 lists SHRG rules for this example. Each row pairs two rules — a CFG rule $A \rightarrow \beta$ for rewriting syntax and an HRG rule $A \rightarrow G$ for rewriting semantics. They are linked through the shared left-hand side label A. Using these rules, syntactic and semantic representations of *visit often Paris* can be constructed synchronously (see Figure 4). For the syntax, the derivation begins with the terminals. Production rules of the form $A \rightarrow \beta$ are applied iteratively to derive non-terminals until the start symbol is reached. Correspondingly, semantics is derived by using a parallel HRG rule $A \rightarrow G$ which shares the left-hand side label with the CFG rule. A hyperedge e in G is rewritten if e has a label n which is a non-terminal symbol of C. In this operation, e is removed from G, and a copy of the previously derived hypergraph H, whose left-hand side label is n, is inserted into G. The external nodes X of H are fused with the nodes connected by e, while other hyperedges in G remain unchanged.

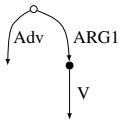
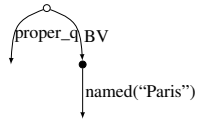
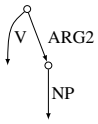
Idx	CFG	HRG	
1	V → visit	V →	
2	Adv → often	Adv →	
3	V → V+Adv	V →	
4	NP → Paris	NP →	
5	VP → V+NP	VP →	

Table 1: Induced SHRG rules for *visit often Paris*. Filled nodes in HRG are external nodes.

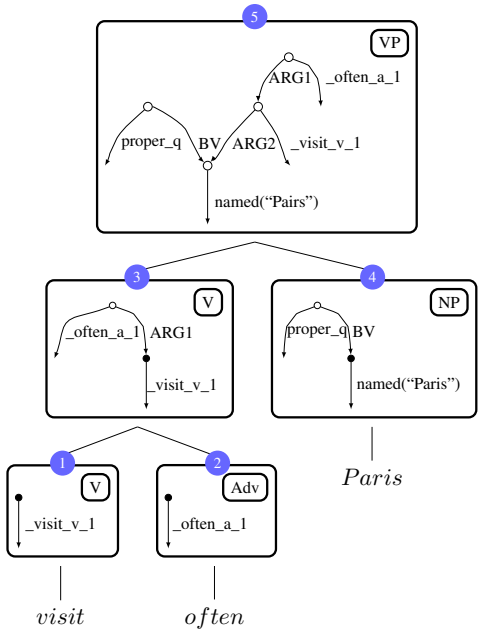


Figure 4: Synchronous construction of syntax and semantics for *visit often Paris*.

2.4 Comparison

We argue that compared to the lexicalist approach, the constructivist framework offers greater flexibility and efficiency in addressing the challenges that "errors" pose to the syntax–semantics interface of ESFL. Taking 1c as an example, within a constructivist view, its transposition "error" can be accounted for by a constructional schema, where the CFG rule allows for post-verbal adverb attachment, while the corresponding HRG rule ensures that the adverb takes the verb as its first argument. This solution is straightforward as it simply adapts existing rules rather than introducing new ones. In contrast, lexicalist solutions such as Combinatory Categorial Grammar (CCG; Steedman, 1987, 2000) seems more inflexible and complex (see Figure 5). It treats lexical items — each associated with a set of syntactico-semantic features — as the fundamental building blocks of language, so the "error" could only be resolved by introducing a new lexical entry for *often*. Moreover, compared to the constructivist approach which allows for the synchronous construction of syntax–semantics representations, this lexicalist approach requires prior syntactic application for semantic derivation, which may decrease efficiency in modeling ESFL.

3 ESFL SemBank

We apply the SHRG-based syntax–semantics interface to ESFL data, resulting a syntactico-semantic

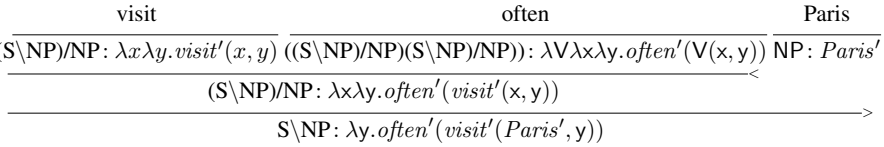


Figure 5: a lexicalist solution in CCG-style. To solve the transposition error, *often*, is assigned a new category.

ESFL SemBank. Generally speaking, its development involves three main steps:

- **Selection of Holistic Graphs:** silver semantic graphs from Zhao et al. (2020) are manually labeled as accepted or rejected.
- **Extraction and Modification of Composition Rules:** accepted graphs are stored for the extraction of SHRG rules, while rejected ones undergo further refinement.
- **Selection of Composition Rules:** semantics of unparsable ESFL sentences in Zhao et al. (2020) is derived through manual selection of composition rules.

3.1 Selection of Holistic Graphs

We start by manually categorizing holistic ESFL semantic graphs, silver outputs from Zhao et al. (2020)[2], into *accept* and *reject* ones. This process ensures a reliable foundation for our sembank and facilitates the creation of a core inventory of reusable rules. The annotation is detailed below.

Materials The original annotated data is drawn from the FCE dataset (Yannakoudakis et al., 2011), which includes essays written by upper-intermediate English learners, along with corrections by native English speakers. Building on the FCE data, Berzak et al. (2016) provide annotations of part-of-speech (POS) tags and syntactic dependencies within the Universal Dependencies (Nivre et al., 2016; de Marneffe et al., 2014) framework for both ESFL sentences and corrected ones (CESFL), resulting in the Treebank of Learner English (TLE). Extending this work, Zhao et al. (2020) create a semantic resource: they parse both ESFL and CESFL sentences in the FCE dataset using an ERG-based processor, ACE parser[3] and apply a reranking model to automatically select the most accurate parse based on alignment with TLE's syntactic annotations. These silver-standard semantic graphs then serve as the material for our manual selection process.

Procedure To begin, a team consisting of one Ph.D. student and two undergraduate majoring in linguistics undertakes training in English Resource Semantics (ERS; Flickinger et al., 2014b). ERS is a semantic framework derived from the English Resource Grammar (ERG; Flickinger, 2000) within Head-driven Phrase Structure Grammar (HPSG; Pollard and Sag, 1994). This training is essential because the previously parsed semantic graphs are represented using this framework.

Subsequently, the annotators use DeepBank (Flickinger et al., 2012), an ERS annotation resource that effectively handles diverse linguistic phenomena in English and accurately represents their semantic interpretations, as a reference. They assess whether the graphs provided by Zhao et al. (2020) accurately capture semantics of ESFL. For each graph, they assign labels such as *accept*, *reject* and *abandon*. The *abandon* label is used when the accuracy of a semantic graph cannot be determined because of insufficient context.

Once the three annotators achieve a high level of inter-annotator agreement (IAA) through several rounds of training, comparison and discussion, the process becomes more efficient: some of the remaining instances are randomly assigned to two annotators for cross-validation, while others are annotated by a single annotator.

Quality This paper utilizes percentage inter-annotator agreement (IAA) to assess the quality of our annotations. As shown in Table 2, our three annotators achieve an exceptionally high IAA, exceeding 99%. For the sentences that are double-annotated, the IAAs fluctuate between 97% and 99%. These high-level agreements indicate a strong consensus among our annotators regarding the semantic interpretation of the ESFL sentences represented within the ERS framework.

[2]Zhao et al. (2020) parsed ESFL sentences using an English-based parser, with outputs reranked based on alignment with syntactic annotations from the Treebank of Learner English (TLE; Berzak et al., 2016).

[3]https://sweaglesw.org/linguistics/ace/

	ESFL	CEFSL
Anno1-Anno2-Anno3	99.29	99.48
Anno1-Anno2	98.95	98.19
Anno1-Anno3	97.82	98.77
Anno2-Anno2	98.28	99.12

Table 2: Inter-annotator agreements (IAAs) of triple-annotated or double-annotated sentences.

Result We annotate 1567 ESFL and 2189 CESFL sentences in total. After excluding records with inconsistent or *abandon* tags, 1543 ESFL and 2138 CESFL sentences remain valid. Table 3 shows the label distribution for these valid sentences: 46.92% of ESFL sentences are accepted and 53.08% rejected, while 61.04% of CESFL sentences are accepted and 38.96% rejected.

	ESL		CESL	
	#num	#per	#num	#per
Triple-acc	64	45.71%	129	66.84%
Triple-rej	76	54.28%	64	33.16%
Triple-all	140	100.00%	193	100.00%
Double-acc	586	46.03%	1051	59.65%
Double-rej	687	53.97%	711	40.35%
Double-all	1273	100.00%	1762	100.00%
Single-acc	74	56.92%	125	68.31%
Single-rej	56	43.08%	58	31.69%
Single-all	130	100.00%	183	100.00%
Overall-acc	724	46.92%	1305	61.04%
Overall-rej	819	53.08%	833	38.96%
Overall-all	1543	100.00%	2138	100.00%

Table 3: Numbers (#num) and percentages (#per) of valid sentences whose semantic graphs accepted (acc) or rejected (rej) by three, two, or one annotators.

3.2 Extraction and Modification of Composition Rules

We decompose all graphs into compositional constructions represented by SHRG rules (§3.2.1) and modify a few problematic components in rejected graphs to "salvage" them (§3.2.2).

3.2.1 Extraction

Pre-processing Before extracting SHRG rules, we reference the parsing results from the ACE parser in Zhao et al. (2020) to retrieve phrase structures in the Penn Treebank (PTB) style.

As shown in Figure 6, even though the retrieved trees, compared to the original HPSG derivation trees, involve only the renaming of non-terminal nodes without altering the structural configuration, the pre-processing is of great importance. This is because in a lexicalist framework like HPSG, derivation trees are not themselves objects of grammatical description; their nodes and edges merely indicate which constituents are combined and how their syntactico-semantic information is composed. Constituents, each associated with a syntactico-semantic feature structure, serve as the fundamental building blocks of grammatical description. However, this processing style contrasts with the constructivist perspective, which asserts that grammatical schemata are an essential component of linguistic competence. From this viewpoint, adopting a more general and coarse-grained PTB-style node set becomes essential. Within this context, trees can be genuinely regarded as a component of grammar, representing an interconnected system of systematically learnable symbolic mappings between forms and meanings.

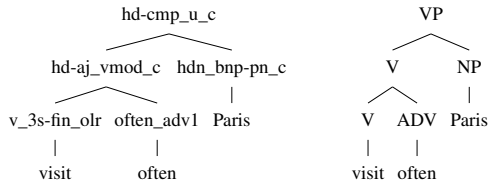


Figure 6: An HPSG derivation tree and its the corresponding PTB-style syntactic analysis.

Algorithm This paper largely follows the extraction algorithm proposed by Chen et al. (2018) to induce SHRG rules. As illustrated in Figure 7, SHRG rules are extracted iteratively by post-order traversal of the tree, linking each semantic subgraph to its corresponding syntactic tree based on shared span information from the surface string. We further distinguish nodes in the identified subgraph at each step: nodes with all connected hyperedges within the subgraph are classified as internal, while others are external.

We argue that the extraction algorithm successfully achieves both comprehensiveness and flexibility. On the one hand, the algorithm ensures comprehensiveness by accounting for all elements: every constituent of the syntactic tree is either mapped to the overall semantic representation or explicitly identified as semantically vacuous. On the other hand, we adapt the original algorithm to allow edges in the semantic graphs not to exactly align with the spans of the syntactic tree nodes,

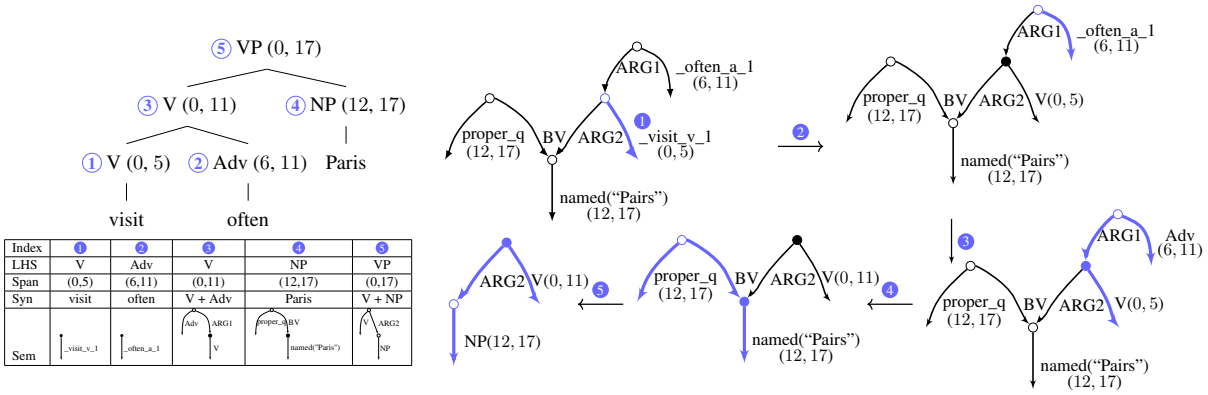


Figure 7: Extraction of SHRG rules for *visit often Paris* using post-order tree traversal. Edges in semantic graphs are identified through span alignment with syntactic nodes, forming subgraphs (colored). External nodes (solid) have connected edges outside the subgraph. LHS denotes shared left-hand side labels.

which enhance its flexibility for ESFL.

Inventory of SHRG Rules We apply the extraction algorithm to all ESFL and CESFL sentences, as well as to DeepBank, a resource of native English sentences. The SHRG rules induced from accepted sentences and DeepBank form a reference inventory, from which rules can be selected to modify the ill-formed fragments in decomposed rejected graphs. Table 4 presents the SHRG rule inventory induced from various datasets.

Data Source	#original	#non-lexical
ESFL-accept	3538	1283
CESFL-accept	7185	3201
DeepBank	44187	20524

Table 4: Induced SHRG rules by dataset: original (total rules) and non-lexical (rules not for a single).

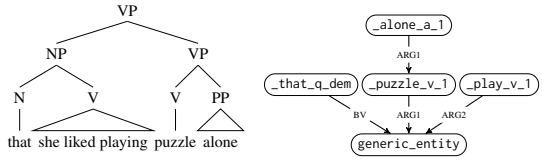
3.2.2 Modification

Building on the induced SHRG rules, we manually refine the composition rules for rejected ESFL semantic graphs. Broadly, we argue that the wide variety of form–meaning mappings in ESFL can be grouped into three categories, each of which can be effectively addressed by our constructivist approach.

Diversity of Grammatical Forms ESFL employs diverse grammatical forms to express grammatical meanings such as Number, Case, Tense, Aspect, Voice, and Agreement. It uses not only morphological markers typical of English, but also zero forms, word order, and functional words.

Consider Sentence 2 as an example. In this case, number — typically marked by the morphological suffix *-s* in native English — is expressed by a zero form in ESFL. This frequently leads to parsing errors with the ACE parser, resulting in inaccurate analysis as shown by Figure 8.

(2) *I was impressed when I heard that she liked playing* ***puzzle*** *alone.*



(a) Original Syn (b) Original Sem

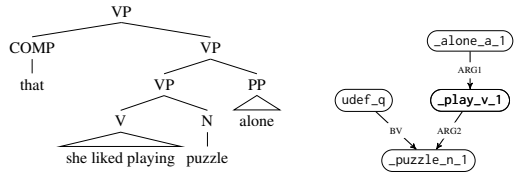
Figure 8: Relevant parts of original syntactic trees and semantic graphs for Sentence 2.

LHS	N_{ori}	V_{ori}	$COMP_{mod}$	N_{mod}
Syn	that	puzzle	that	puzzle
Sem	_that_q_dem, BV, generic_entity	_puzzle_v_1	∅	udef_q, BV, _puzzle_n_1

Table 5: Original and modified rules for sentence 2.

Since this phenomenon involves determiners and often lacks clear markers of number, we address it using an SHRG rule with the abstract predicate `udef_q`, along with other relevant rules (see Table 5). These rules are then syntactically recombined to yield accurate semantic representations (see Figure 9). More broadly, adapting existing rules to accommodate such grammatical forms is a general strategy we employ.

Diversity of Syntactic Derivations ESFL may impose unique constraints on phenomena involving



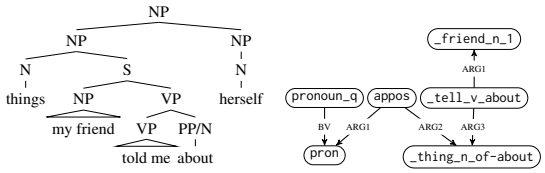
(a) Modified Syn (b) Modified Sem

Figure 9: Relevant parts of modified syntactic trees and semantic graphs for Sentence 2.

syntactic derivations, such as Binding, Ellipsis, and Filler-Gap constructions. Unlike generative grammar, which analyzes these phenomena as results of deep-structure operations like binding, deletion, or movement, our framework accounts for them through novel syntax–semantics mappings.

We illustrate this type of grammatical variation in ESFL with Sentence 3, which involves binding — a phenomenon concerning with constraints on anaphoric expressions such as pronouns and reflexives. According to binding theory (Chomsky, 1981, 1986), a reflexive in native English must be bound by the the noun within its local domain. However, this constraint is more loosely applied in ESFL. In this example, the reflexive *herself* is interpreted as referring to *Pat*, rather than the grammatically appropriate antecedent *my friends*, thus violating standard binding principles and leading to problematic analysis (see Figure 10).

(3) *At first, Pat denied all the things my friends told me about* ***herself*** *but she finally agreed with that.*



(a) Original Syn (b) Original Sem

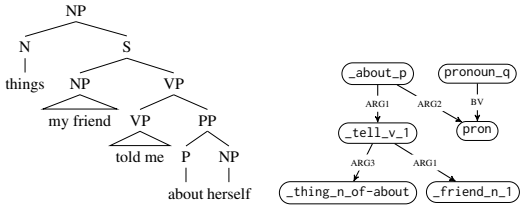
Figure 10: Relevant parts of original syntactic trees and semantic graphs for Sentence 3.

We utilize the SHRG rules detailed in Table 6 to resolve this discrepancy, and derive appropriate semantics as shown by Figure 11.

Diversity of Lexical Items ESFL also displays notable lexical variation, especially in verbs, whose argument structures — which govern both the number and realization of arguments — often diverge from those in standard English.

LHS	PP/N$_{ori}$	NP$_{ori}$	P$_{mod}$	PP$_{mod}$
Syn	about	NP + NP	about	P + NP
Sem	∅	appos ARG1 ARG2 NP NP	_about_p	P ARG2 NP

Table 6: Original and modified rules for example 3.

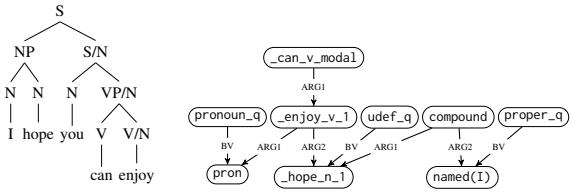


(a) Modified Syn (b) Modified Sem

Figure 11: Relevant parts of modified syntactic trees and semantic graphs for sentence 3.

For example, in Sentence 4, the verb enjoy lacks its required object, resulting in a different argument structure from that expected in native English. Such variations challenge the ACE parser, leading to inaccurate representations shown in Figure 12.

(4) *I hope you can* ***enjoy****.*



(a) Original Syn (b) Original Sem

Figure 12: Relevant parts of original syntactic trees and semantic graphs for Sentence 4.

To handle these discrepancies, we adapt or extend existing SHRG rules. Table 7 lists both the original and modified rules, and Figure 13 further shows the correct syntactic and semantic analysis after modifying SHRG rules correspondingly.

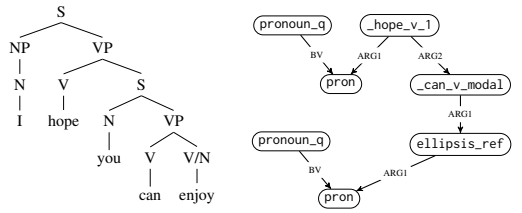
3.3 Selection of Composition Rules

The following details how we manually select composition rules from the SHRG inventory to generate syntactico-semantic representations for previously unparsable sentences.

Material Zhao et al. (2020) reports that because of limit of the ACE parser and inconsistent tokenization, it successfully processes 52.50% ESFL

LHS	N_{ori}	NP_{ori}	V/N_{ori}	S_{ori}
Syn	hope	N + N	enjoy	NP + S/N
Sem	_hope_n_1	compound ARG1 ARG2 N N	_enjoy_v_1	S/N ARG2 NP
LHS	V_{mod}	N_{mod}	V/N_{mod}	VP_{mod}
Syn	hope	I	enjoy	V + S
Sem	_hope_v_1	pronoun_q BV pron	ellipsis_ref	V ARG2 S

Table 7: Original and modified rules for example 4.



(a) Modified Syn (b) Modified Sem

Figure 13: Relevant parts of modified syntactic trees and semantic graphs for Sentence 4.

sentences, leaving 2494 ones unanalyzed. We randomly select 100 sentences from them as an initial investigation. Although limited in scale, it is methodologically justified, as the annotation process for these unparsable sentences — requiring manual construction of syntactic trees followed by the selection of SHRG rules — is similar to that of the majority of rejected cases. Statistical analysis shows that 81.2% (665 graphs) of the rejected instances involve reconstruction of syntactic trees.

Process The annotation process consists of two phases. In the first one, an annotator constructs semantic graphs by selecting SHRG rules, while the other reviews these graphs to assess their acceptability. In the second phase, both annotators independently annotate the remaining 50 sentences.

Result The annotation results from both phases are shown in Table 8. For the first 50 sentences, we calculate the proportion of accepted graphs to assess the level of agreement between the two annotators. In the second phase, we calculate the average S-match score for the 50 independently annotated graphs as a element-wise measure of consistency.

3.4 Summary

As Table 9 presents, following the three steps above, we develop an ESFL SemBank, which contains manually annotated syntactico-semantic representations of 1643 ESFL sentences.

	First Phase	Second Phase
Consistency Score	94%	90.93

Table 8: Consistency scores in selecting rules.

Step	Manner	Number
1	Manually Accepted	724
2	Manually Modified	819
3	Manually Composed	100
		Overall: 1643

Table 9: Development of ESFL SemBank at each step.

4 Conclusion

Compositionality is not only a key notion of theoretical linguistics, as linguistics concerns itself with understanding the human language capacity, but also an emerging crucial topic in cognitive science and artificial intelligence (Frankland and Greene, 2020; Hampton and Winter, 2017; Hupkes et al., 2020) given the defining role played by compositionality in distinguishing human language from animal communication systems. However, the "erroneous" tokens and grammars frequently observed in non-native language varieties, particularly ESFL, present significant challenges to systematic semantic derivation and seemingly undermine the principle of compositionality.

This paper addresses the issue by integrating insights from constructivist theories and validating this approach through computational modeling. Our findings show that, despite surface-level syntactic imperfections, compositionality operates similarly in the underlying mechanisms of both ESFL and native English. We also present a more accurate, consistent, and comprehensive syntactico-semantic bank, containing robust and syntactically informed annotations for 1643 ESFL sentences.

Given the global prevalence of ESFL, we believe our contributions, including the proposed framework and the SemBank, are of substantial importance. By providing a rich empirical foundation, they could address long-standing questions and open new avenues for both theoretical and computational linguistics, as well as other fields beyond.

Limitations

While the syntactico-semantic bank for ESFL has been developed, the scalability of manual annotation through automatic or semi-automatic methods remains an open question that requires further investigation. Future work could focus on the semantic parsing for ESFL to further enhance the generalizability of the proposed framework.

Acknowledgments

We express our gratitude for the assistance and support rendered by annotators Yutong Zhang, Yixuan Wang, and Huangyang Xiao. Additionally, our heartfelt thanks go to all the reviewers and editors for their contributions.

References

Christian Adjemian. 1976. On the nature of interlanguage systems. Language learning, 26(2):297–320.

Gustavo Aguilar and Thamar Solorio. 2020. From English to code-switching: Transfer learning with strong morphological clues. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8033–8044, Online. Association for Computational Linguistics.

Elizabeth Bates and Brian MacWhinney. 1987. Competition, variation, and language learning. In Elizabeth Bates, editor, Mechanisms of language acquisition, pages 157–193. Erlbaum, Hillsdale, NJ.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for learner English. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 737–746, Berlin, Germany. Association for Computational Linguistics.

Douglas Biber and Randi Reppen. 2015. The Cambridge handbook of English corpus linguistics. Cambridge University Press.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. Computational Linguistics, pages 643–701.

Yufei Chen and Weiwei Sun. 2020. Parsing into variable-in-situ logico-semantic graphs. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6772–6782, Online. Association for Computational Linguistics.

Yufei Chen, Weiwei Sun, and Xiaojun Wan. 2018. Accurate SHRG-based semantic parsing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 408–418, Melbourne, Australia. Association for Computational Linguistics.

Noam Chomsky. 1981. Lectures on Government and Binding: The Pisa Lectures. Mouton de Gruyter.

Noam Chomsky. 1986. Knowledge of Language: Its Nature, Origin, and Use. Prager.

Harald Clahsen and Claudia Felser. 2006. How native-like is non-native language processing? Trends in cognitive sciences, 10(12):564–570.

William Croft. 2001. Radical construction grammar: Syntactic theory in typological perspective. Oxford University Press.

William Croft and D. Alan Cruse. 2004. Cognitive linguistics. Cambridge Textbooks in Linguistics. Cambridge University Press.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. Ethnologue: Languages of the World, twenty-seventh edition. SIL International, Dallas, Texas. Online version: http://www.ethnologue.com.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. Natural Language Engineering, 6(1):15–28.

Dan Flickinger, Emily M. Bender, and Stephan Oepen. 2014a. Towards an encyclopedia of compositional semantics: Documenting the interface of the English Resource Grammar. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), pages 875–881. European Language Resources Association (ELRA).

Dan Flickinger, Emily M. Bender, and Stephan Oepen. 2014b. Towards an encyclopedia of compositional semantics: Documenting the interface of the English Resource Grammar. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 875–881, Reykjavik, Iceland. European Language Resources Association (ELRA).

Dan Flickinger, Yi Zhang, and Valia Kordoni. 2012. Deepbank: A dynamically annotated treebank of the wall street journal. In Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories, pages 85–96.

Steven M Frankland and Joshua D Greene. 2020. Concepts and compositionality: in search of the brain's language of thought. Annual review of psychology, 71(1):273–303.

Jeroen Geertzen, Theodora Alexopoulou, Anna Korhonen, et al. 2013. Automatic linguistic annotation of

large scale l2 databases: The ef-cambridge open language database (efcamdat). In Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project, pages 240–254.

Adele E Goldberg. 1995. Constructions: A construction grammar approach to argument structure. Chicago UP.

Adele E Goldberg. 2003. Constructions: A new theoretical approach to language. Trends in cognitive sciences, 7(5):219–224.

Adele E Goldberg. 2013. Argument structure constructions versus lexical rules or derivational verb templates. Mind & Language, 28(4):435–465.

Sylviane Granger. 2003. The international corpus of learner english: a new resource for foreign language learning and teaching and second language acquisition research. Tesol Quarterly, 37(3):538–546.

Sylviane Granger. 2014. Learner English on computer. Routledge.

James A Hampton and Yoad Winter. 2017. Compositionality and concepts in linguistics and psychology. Springer Nature.

Thomas Hoffmann and Graeme Trousdale. 2013. The Oxford handbook of construction grammar. Oxford University Press.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? Journal of Artificial Intelligence Research, 67:757–795.

Ray Jackendoff. 2000. The representational structures of the language faculty and their interactions. In The Neurocognition of Language. Oxford University Press.

Ray Jackendoff. 2002. Foundations of Language: Brain, Meaning, Grammar, Evolution. Oxford University Press.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3575–3585, Online. Association for Computational Linguistics.

Ingrid Kurz. 2008. The impact of non-native english on students' interpreting performance. In Efforts and models in interpreting and translation research: A tribute to Daniel Gile, pages 179–192. Benjamins Publishing.

Ronald W Langacker. 1987. Nouns and verbs. Language, pages 53–94.

Brian MacWhinney. 2000. The CHILDES project: The database, volume 2. Psychology Press.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Barbara Partee. 1984. Compositionality. In Varieties of Formal Semantics: Proceedings of the 4th Amsterdam Colloquium, volume 3, pages 281–311. Foris Dordrecht.

Carl Pollard and Ivan A Sag. 1994. Head-driven phrase structure grammar. University of Chicago Press.

Peter Robinson and Nick C Ellis. 2008. Handbook of cognitive linguistics and second language acquisition. Routledge, London.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2007. High-accuracy annotation and parsing of CHILDES transcripts. In Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition, pages 25–32, Prague, Czech Republic. Association for Computational Linguistics.

Mark Steedman. 1987. Combinatory grammars and parasitic gaps. Natural Language & Linguistic Theory, 5(3):403–439.

Mark Steedman. 2000. The Syntactic Process. MIT Press, Cambridge, MA.

Michael Tomasello. 2003. Constructing a Language: A Usage-Based Theory of Language Acquisition. Harvard University Press.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Yajie Ye and Weiwei Sun. 2020. Exact yet efficient graph parsing, bi-directional locality and the constructivist hypothesis. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4100–4110, Online. Association for Computational Linguistics.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9220–9230, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuanyuan Zhao, Weiwei Sun, Junjie Cao, and Xiaojun Wan. 2020. Semantic parsing for English as a second language. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6783–6794, Online. Association for Computational Linguistics.