Subcategorization frame identification for learner English

Huang, Yan; Murakami, Akira; Alexopoulou, Theodora; Korhonen, Anna

DOI:
10.1075/ijcl.18097.hua

Document Version
Peer reviewed version

Citation for published version (Harvard):
Huang, Y, Murakami, A, Alexopoulou, T & Korhonen, A 2021, 'Subcategorization frame identification for learner English', *International Journal of Corpus Linguistics*, vol. 26, no. 2, pp. 187-218.
https://doi.org/10.1075/ijcl.18097.hua

Subcategorization frame identification for learner English

Yan Huang [i], Akira Murakami [ii], Theodora Alexopoulou [i] & Anna Korhonen [i]

[i] University of Cambridge | [ii] University of Birmingham

As large-scale learner corpora become increasingly available, it is vital that natural language processing (NLP) technology is developed to provide rich linguistic annotations necessary for second language (L2) research. We present a system for automatically analyzing subcategorization frames (SCFs) for learner English. SCFs link lexis with morphosyntax, shedding light to the interplay between lexical and structural information in learner language. Meanwhile, SCFs are crucial to the study of a wide range of phenomena including individual verbs, verb classes and varying syntactic structures. To illustrate the usefulness of our system for learner corpus research and second language acquisition (SLA), we investigate how L2 learners diversify their use of SCFs in text and how this diversity changes with L2 proficiency.

Keywords: subcategorization, verb-argument construction, SCF identification, second language acquisition, natural language processing

1. Introduction

Recent decades have seen emergence of increasingly large learner corpora. Such corpora provide exciting opportunities for second language (L2) research. They can help to improve the empirical scope and robustness of claims, and can support the discovery of linguistic phenomena that have previously evaded human intuition. To fully exploit the power of learner corpora, it is essential to analyze syntactic features. Syntactic analysis provides not only indices to syntactic phenomena in corpora, but also clues for retrieving lexical, semantic and pragmatic information (Meurers et al., 2013).

As learner corpora become larger, manual annotation of syntactic information becomes less feasible, and automatic techniques can be beneficial or even necessary.

Corpus linguists have used part-of-speech (POS) taggers and parsers to extract syntactic patterns from large corpora (Gries & Berez, 2017). Based on these patterns, researchers can analyze abstract syntactic features, such as syntactic structural similarity (Graesser et al., 2011) and syntactic complexity (Biber, 1988; Chen & Meurers, 2019; Kyle, 2016; Lu, 2010) of learner English efficiently.

However, some specific syntactic phenomena of interest to L2 research cannot be easily investigated by existing syntactic analysis systems. Subcategorization is one such phenomenon. Subcategorization specifies the syntactic context in which a word of a particular category may appear. More specifically, a subcategorization frame (SCF) defines the number and types of syntactic complements required by a predicate (Chomsky, 1965). SCF is the linguistic realization of argument structure which is central to all grammar theories (Jackendoff, n.d.). While SCF originates from constituency grammar, alternative theories approach the phenomenon as valency (Tesnière, 1965) or construction (Goldberg, 1995).

SCF is interesting for corpus linguistics. Researchers have used learner corpora to investigate L2 acquisition of SCFs. They have studied how learner SCFs develop over time (Ellis et al., 2016; Tono, 2004), and how various factors such as input (Ellis et al., 2016) and verb semantics (Ellis et al., 2016; Römer et al., 2014, 2015) affect that development. These studies not only have practical implications for L2 education, but also allow researchers to provide empirical insights into human linguistic capacity and cognitive mechanisms. For example, by analyzing the relation between the frequency of a verbal construction and its acquisition, Ellis et al. (2016) testified the usage-based theory about constructional learning. Meanwhile, many interesting questions about L2 acquisition of SCFs remain uninvestigated, e.g. whether there is difference between L1 and L2 acquisition of SCFs, and how L1 background affects L2 acquisition of SCFs.

Furthermore, SCF is a morphosyntactic structure which can serve as the basis of various corpus analyses. Take linguistic complexity as an example. Complexity, along with accuracy and fluency (CAF), has become a principal angle for defining or characterizing language proficiency (Norris & Ortega, 2009). Corpus linguists have conducted numerous studies on L2 linguistic complexity. A fundamental issue in this area concerns how to operationalize and measure linguistic complexity. Linguistic complexity is commonly defined as the ability to use a wide range of sophisticated linguistic units (Bulté & Housen, 2012; Wolfe-Quintero et al., 1998). This means that

the diversity of the linguistic units is an important factor. Existing linguistic diversity measures have considered the linguistic units of words, phrases, clauses and sentences, but not SCFs. For example, the Mean Length of Clause (MLC) measures the average number of words per clause (Lu, 2010), and Coordinate Phrases per Clause (CP/C) measures the number of coordinate phrases per clause (Kyle, 2016). Prior research showed that different diversity measures may have different relationships with proficiency level, gauging different dimensions of linguistic complexity (Norris & Ortega, 2009). Due to the unique morphosyntactic nature of SCF, it is interesting to investigate how the diversity of SCF use is related to language proficiency. SCF-based linguistic diversity measures may provide a new angle to linguistic complexity research.

Despite the large potential of SCF for corpus linguistics, research in this field has been limited by the lack of automatic tools to analyze SCF. So far, researchers have extracted SCF information from corpora manually, or semi-manually with a POS tagger (Ellis et al., 2016) or parser (Meurers et al., 2013; Tono, 2004). The syntactic information provided by these general NLP systems is not straight-forward for searching SCFs. Researchers have to define complicated rules to extract potential SCF patterns. The precision and recall of the extraction rules are low, and extensive human effort is required to distinguish arguments and adjuncts. For example, to extract the "V-P" type of SCFs (e.g. *he talked about the weather*) based on POS tags and the dependency labels, Römer et al. (2015) conducted three rounds of search refinement; each round involved the design of search rules, the manual edition of search results (1,500 sentences for each SCF pattern), and the evaluation of search accuracy. After the painstaking effort, they achieved an average of 78% precision, 53% recall and 61% F1 score. The time-consuming effort has restricted the amount of SCF data that has been analyzed in corpus-based studies, consequently limiting the power of corpus-based approaches to this topic.

This paper presents the first SCF identification system for learner English[1]. Specifically, our system can label individual occurrences of verbs in learner corpora for a set of 49 distinct SCFs ranging from basic transitive and intransitive frames to complicated frames that involve prepositional, verbal or clausal complements. To illustrate the usefulness of the system, we use a large-scale learner corpus to

[1] https://github.com/cambridgeltl/subcategorization-frames-and-learner-English-data

investigate how L2 English learners diversify their use of SCFs at different L2 proficiency levels.

Our paper is organized as follows: Section 2 reviews the major challenge in analyzing SCFs and introduces existing schemes and NLP systems for SCFs. Section 3 presents the development of our SCF identification system, while Section 5 illustrates the usefulness of the system by presenting a corpus-based study of L2 SCF acquisition. Section 6 summarizes the findings of our corpus-based study and indicates the potential benefits of our SCF system for linguistic research.

2. Subcategorization frames and their automatic identification

SCF requires a distinction between complements and adjuncts. Complements are expected to complete the meaning of the predicate. For example, the direct object *the pen* and the prepositional object *on the chair* are required by the predicate *put* in (1). Adjuncts, on the other hand, are peripheral. For example, the prepositional object *in a hurry* in (2) is not required by the predicate *walked*. The complement-adjunct distinction is not only important for the theoretical definition of SCF, but also vital to any study of SCF that concerns the relation between predicates and verbs, or concerns the meaning of a SCF construction. This is because complements have a close relation to the predicate, and strongly indicate the meaning of the predicate and the SCF. Contrastingly, adjuncts can be used with many predicates freely, and have no such indication. For example, in (2), the prepositional object *in a hurry* can be used with a wide range of predicates, including the predicates in (1) and (3).

(1) She put [the pen] [on the chair].

(2) She walked (in a hurry).

(3) She can sing [the song].

However, distinguishing complements and adjuncts can be difficult. Numerous criteria have been proposed to distinguish complement from adjuncts (Somers, 1984), but no criteria are applicable to all situations and they sometimes yield inconsistent results. For example, the most common test for distinguishing complements and adjuncts is the elimination test (Helbig & Schenkel, 1991): an element is eliminated

from the sentence; if the remaining sentence is ungrammatical, the element is a complement; otherwise, the element is an adjunct. For example, *on the chair* in (1) is a complement because it cannot be eliminated, whereas *in a hurry* in (2) is an adjunct because it can be eliminated. The elimination test, however, fails in (3). In (3), the nominal object *the song* can be eliminated but is not an adjunct, because *the song* essentially completes the meaning of the predicate. Furthermore, there is a slight difference between the meanings of the predicate *sing* with and without *the song*: the former refers to the ability of singing a particular song, whereas the latter refers to the general ability of "singing well". This illustrates that a complement can be optional, and its presence or absence can affect the meaning of the predicate. The difficulty in distinguishing complements and adjuncts lies in their ambiguous boundaries. Somers (1984) argues that they are prototypes on a spectrum, where intermediate cases and more extreme cases on both ends can be found.

Despite this difficulty, Meyers et al. (1996) summarize a set of sufficient conditions and rules-of-thumb for distinguishing English complements and adjuncts. They empirically prove that these rules are useful for achieving consistent annotation; on average, 91% of the complements classified by an annotator were classified the same by three other annotators.

Computational linguists have developed a number of SCF schemes when constructing large-scale lexicons for real-world NLP applications. Representational works include the manually constructed computational lexicon Comlex (Grishman et al., 1994), which has 92 types of SCF for verbs, and the Alvey NLP Tools dictionary (ANLT) (Boguraev & Briscoe, 1987), which was manually adapted from the electronic version of the Longman Dictionary of Contemporary English (Procter, 1978). By merging and supplementing the SCF schemes of Comlex and ANLT, Briscoe and Carroll (1997) developed a detailed scheme of 163 SCF types for verbs. This scheme was subsequently extended to 168 SCF types (Preiss et al., 2007). In this study, we adopt the scheme of Preiss et al. (2007), the most comprehensive one, as the basis for producing our SCF inventory.

Many NLP systems have been developed to infer the likelihood of SCFs for a verb form in native English (e.g. Briscoe & Carroll, 1997). To illustrate, such a system may infer that the probability of the verb form *put* used with the SCF comprising a direct object and an adverb (e.g. *put the book here*) is 24% (hypothetical). However, these systems cannot annotate SCFs for individual verb tokens, such as *put* in the

particular context of (1), which is the type of linguistic annotation needed in L2 research.

There are only two NLP systems which can annotate SCFs for verb tokens. Baker et al. (2014) proposed an unsupervised method that can cluster verb tokens according to their syntactic context. The clusters are regarded as SCFs. However, the labels of the clusters are unknown, which is uninformative for L2 research. Dušek et al. (2014) developed a system that can assign SCF labels to verb tokens. However, the system applies to only a limited number of verbs, because the system uses a separate SCF classification model for each verb lemma in the training data.

To fill the gap, we propose a SCF identification system that can label the SCFs of individual verb tokens contextualized in sentences. Our system is applicable to any verb token, as we use a unified machine learning model for all verbs. We trained the system on learner English data so that it can identify learner SCF patterns.

3. A SCF identification system for learner English

We approach SCF identification as a supervised classification task, training a classifier on SCF corpora. The following sections introduce our data, method and the evaluation of the system.

3.1 Data

We used learner English and native English datasets. The purpose of including a native English dataset was to increase the training data; since the learner English training data may not include all SCF types, adding native English training data can improve the generalizability of the model for unseen learner English data.

For native English, we adopted a domain-general SCF dataset (Quochi et al., 2014) which contains 6,133 sentences (186,534 word tokens) sampled from the British National Corpus (BNC) (Aston & Burnard, 1998). In each sentence, only one verb was annotated for SCF. The dataset was annotated by a linguist with the fine-grained SCF schemes of (Preiss et al., 2007). We mapped the SCFs to a coarser-grained scheme defined on the popular Stanford typed dependencies (De

Marneffe and Manning 2008). We reduced the granularity for three reasons: first, the distribution of SCFs is Zipfian and many fine-grained SCFs rarely appear in real-world data. Second, learners tend to use simple SCFs, and a coarse-grained scheme provides appropriate granularity for analyzing learner SCFs. Third, a coarse-grained SCF scheme provides a suitable level of specificity for downstream NLP tasks. Our final SCF scheme contained 70 SCF types (See Appendix).

The SCF types were named by the complements involved. The complements can be classified to eight types according to their dependency relations with the predicate: adjectival complement ("acomp"), adverbial modifier ("advmod"), clausal complement ("ccomp"), direct object ("dobj"), indirect object ("iobj"), prepositional object or complement ("prep"), particle ("prt"), and open-clausal complement ("xcomp") [2] . Multiple complements were joined by colons. For example, "dobj_N:iobj" had two complements: "dobj_N" and "iobj". Meanwhile, "_" denoted the POS of the head word of a complement or whether the complement was introduced by a wh-word. For example, "ccomp_VTENSED" meant that the head word of the clausal complement was a finite verb. Moreover, "=>" denoted the dependent of a complement, and the dependent may be lexicalized and denoted by "-". For example, "ccomp_VTENSED=>mark-that" (e.g. *It indicated [that he left]*) meant that the clausal complement has a dependent of marker (a word that introduced a subordinate finite clause), and the marker is lexicalized by *that*.

We used learner data from the EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen et al., 2013). EFCAMDAT contains writings submitted to Englishtown, the online school of Education First. At the time of our experiments, EFCAMDAT had 44,090,870 words written by 109,569 learners. The writings covered 128 topics and various writing types such as narrative (e.g. writing a movie plot) and descriptive (e.g. describing your house). The writings spanned 16 proficiency levels covering the whole spectrum A1-C2 of Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). The proficiency levels were allocated to learners after a placement test when they started a course at EF or through successful progression through coursework. There was considerable diversity in learner backgrounds; Brazilian was the most dominant group (35% of the writings), followed by Chinese (21%), Mexican (7%), Russian (7%),

[2] See Appendix for examples of the complements.

German (5%), French (4%) and Italian (4%). The wide range of proficiency levels and nationalities made EFCAMDAT an appropriate data source for the development and testing of our SCF identification system for learner English.

We annotated SCFs manually for a subset of 1,000 sentences (12,003 word tokens) from EFCAMDAT. This subset, hereafter referred to as EF1000, was previously used to evaluate parsers on learner English (Geertzen et al., 2013; Huang et al., 2018). The sentences were randomly sampled with equal representation from each proficiency level and each of the five most represented nationalities (i.e. Chinese, Russian, Brazilian, German, and Italian). EF1000 comes with manual annotations of Penn Treebank POS tags (Marcus et al., 1993) and Stanford typed dependency structure (De Marneffe & Manning, 2008). Figure 1 shows the POS tags and dependency structure of an example sentence *He smiled and thought about whether he should go*. For example, the first word *he* is a pronoun (PRP), serving as the nominal subject ("nsubj") of the verb *smiled*. The Stanford typed dependency scheme is semantics-oriented and treats the verb of a subordinate phrase rather than the subordinating conjunction as the head, e.g. *whether* is a dependent of *go* rather than *about*. EF1000 was also manually annotated for learner errors following the error scheme of Cambridge Learner Corpus (CLC-FCE) (Nicholls, 2003).

INSERT FIGURE 1 HERE

Figure 1. The POS tags and dependency structure of an example sentence

We identified 1,987 verbs from the learner dataset, choosing the verbs according to POS tags and dependency relations:

i. The POS tag of the word contained "VB";

ii. The dependency relation of the word was not "aux" (auxiliary, e.g. *has left*), "auxpass" (passive auxiliary, e.g. *has been*), "amod" (adjectival modifier, e.g. *frozen food*), or "nn" (noun compound modifier, e.g. *the swimming pool*).

We used the SCF inventory of native English to annotate the learner data. In the presence of learner errors, SCFs were annotated based on surface evidence. For example, in the sentence *I waited John*, the SCF of *waited* was annotated as "dobj_N" (a direct object). For a learner SCF that was not in the SCF inventory (e.g. the SCF of

dream in *I dream about travel around the world* contained a prepositional complement erroneously headed by a base-form verb; this SCF can be termed as a new frame called "pcomp_VBARE"), we annotated it as "new frame".

Two Linguistics PhD students participated in the annotation of SCFs. The annotators first learned the SCF inventory and an annotation guideline developed based on Meyers et al. (1996). The annotators then went through two training sessions. In each session, they annotated 100 verb tokens independently. The first author of this paper also annotated the training sentences. At the end of each training session, the annotators and the author compared their annotations, discussing and resolving disagreement. After the training, the two annotators continued to annotate the remaining 1,787 verb tokens independently. 83.7% of their annotations were completely identical. The relatively low agreement was caused by the inherent difficulty in distinguishing complements and adjuncts, and the difficulty increased for learner English, which has more variable structures and learner errors. The first author then reviewed the disagreements and decided the final annotation. The final annotation showed that the incidence of SCF learner errors was low: 12 (0.6%) verb tokens were annotated as "new frame"; 68 (3.4%) verb tokens had wrong SCFs (e.g. *I waited John* instead of *I waited for John*) and 20 verb tokens (1.0%) had fine-grained errors in the choice of prepositions or particle. Since new frames are rare and varied, they cannot be reliably classified by a machine learning model. We therefore removed the verb tokens annotated as "new frame" from the dataset. As a result, the SCF learner corpus contained 1,966 verb tokens. Even though the dataset was small, as Section 4.3 will show, it was sufficient for training our SCF identification system to achieve an accuracy that was close to the inter-annotator agreement.

The native English dataset had 43 SCF types, while the learner dataset had 38 types. Thirty-two types overlapped. This meant only about half of the SCFs in our inventory actually appeared in both datasets. SCF distributions tend to be Zipfian and the SCF types absent in the data were rare in real-world situations. Since each dataset contained SCF types that were absent in the other dataset, using both datasets as training data can increase the coverage of the SCF types.

3.2 Method

We employed a maximum entropy (MaxEnt) model (Berger et al., 1996) as our classifier. MaxEnt has proved to be useful in automatic syntactic analysis such as POS tagging and parsing (Charniak & Johnson, 2005). In general, the model used the features of a verb token to calculate the probability of each SCF in the inventory, and assigned the SCF of the highest probability score to the verb token.

We used four types of linguistic information to create our features: words, POS tags, dependency relations, and word embeddings. The first three features have proved to be useful in capturing SCF information (Baker et al., 2014). Meanwhile, word embeddings are playing a key role in many recent syntactic NLP systems such as dependency parsers (Andor et al., 2016). A word embedding is a distributional vector representation of a word (Mikolov et al., 2013). Semantically similar words tend to have similar word embeddings. For example, the word embedding of *reply* is more similar to that of *respond* than *stay*. Since semantically similar verbs tend to have similar SCFs (Levin, 1993), the word embeddings of predicates can be useful for identifying SCFs.

More specifically, we extracted the following features for a given predicate:

i. The combinations of the word, POS tag, and dependency relation of a child (i.e. a dependent of the predicate), a grandchild (i.e. a dependent of a child), or a great grandchild (i.e. a dependent of a grandchild) that was *whether*, *if*, or a wh-word. These features were intended to capture the potential complements of the predicate. Take the predicate *thought* in Figure 1 for example. Its child *about*, grandchild *go*, and great grandchild *whether* were considered for feature extraction. The features extracted for the child *about*, of which the POS tag was IN and the dependency relation was "prep", included seven combinations: "ch_about", "ch_IN", "ch_prep", "ch_about_IN", "ch_about_prep", "ch_IN_prep" and "ch_about_IN_prep".

ii. The full combination of the word, POS tag and dependency relation of a parent, a grandparent or a sibling of the predicate. These features were intended to capture information in the head words and conjuncts of the predicate which may be useful for inferring SCF. For example, the feature extracted for the sibling *smiled* was "sb_thought_VBD_root".

iii. The n-grams of the lexicalized or unlexicalized combinations of the word, POS tag, and dependency relation of the neighboring words of the predicate. These

features were intended to capture the context of the predicates. At most one word to the left and three words to the right of the predicate were considered. This imbalanced context window was designed following the observation that most SCF information is located to the right of a predicate. We extracted unigrams and bigrams within the window. The lexicalized features included both words and dependency relations, whereas the unlexicalized ones included the dependency relation and the position of the word with regard to the predicate. For example, the unlexicalized bi-gram feature for the two neighboring words *about* and *whether* was "du_1_prep_2_mark". The word position information is excluded from the lexicalized features to avoid data sparsity for machine learning.

iv. The word and word embedding of the predicate. These features were intended to capture information about the predicate.

Since we intended to develop a SCF identification system that requires no manual syntactic annotation as input, we used SyntaxNet, a state-of-the-art syntactic parser for English (Andor et al., 2016), to extract POS tags and dependency relations. SyntaxNet was trained on Penn Treebank, following the same syntactic schemes for EF1000, i.e. Penn Treebank POS tags and Stanford typed dependency. SyntaxNet achieves an accuracy of 97.4% on POS tagging and 92.8% LAS on dependency parsing for native English data from *Wall Street Journal* (Marcus et al., 1993). We employed a word embedding model trained on the English Polyglot Wikipedia corpus (Al-Rfou' et al., 2013) with a dimensionality of 300. This model has performed well on some semantic NLP tasks (Gerz et al., 2016). Out-of-vocabulary words (i.e. words absent in the training corpus) were mapped to vectors of zeros.

3.3 Training and evaluation

We trained the model in two data settings. The first setting used learner data only. We conducted a 10-fold cross-validation on the learner data. In other words, we partitioned the learner data into 10 subsets; we trained the model on nine subsets at a time and tested the model on the remaining subset; this process was repeated 10 times so that all subsets were used for testing. Cross-validation is a standard method for

model evaluation in machine learning. Since a model is tested on the subset of data which is not included in model training, the method ensures the generalizability of the test results. The average accuracy of our first data setting was 82.1%. In the second setting, we added native data for the training. The average accuracy of this setting was 84.2%. This meant that adding native data during training helped to improve the accuracy of SCF identification on learner data. As a result, we trained our model on both learner and native data. To evaluate the usefulness of the model features, we experimented with a baseline model which used the predicate as the only feature. The accuracy of this model was only 38.0%. Meanwhile, we conducted a leave-one-out experiment with 10-fold cross-validation on the model, i.e. we removed one type of information (words, POS tags, dependency relations or word embedding) from the features at a time during the training. All experiments led to a decreased accuracy, which meant all types of feature were important for SCF identification. As a result, we used the full features to train the model. The final model was regarded as the SCF identification system.

To evaluate how much the SCF identification system performed better than the parser in distinguishing complements and adjuncts, we implemented a rule-based baseline system as follows: for a verb token, the baseline extracted all the dependents which had the potential to be complements (i.e. the eight types of dependency relations in Section 4.1) as a proxy of the SCF, following Meurers et al. (2013) and Kyle (2016). Copula verbs were converted to be the heads of their complements before the extraction. The SCFs were named after the dependency relations only (e.g. both "ccomp_VTENSED" and "ccomp_VTENSED=>mark-that" were named as "ccomp"). Such SCFs correspond to a coarser level of our SCF scheme. The baseline system did not extract finer-grained SCFs because it would require non-trivial effort to design such extraction rules, while a coarse-level evaluation sufficed our evaluation purpose. On this coarse level, the baseline model achieved an accuracy of 51.1%, whereas the SCF identification system achieved 84.9%. This result shows that the SCF identification can better distinguish complements and adjuncts than the parser, improving the accuracy by more than 30%.

In the following sections, we report the accuracy of the system on individual SCFs, and analyze the SCF errors made by the system.

Table 1. Precision (P.), recall (R.) and FI score of SCF identification of individual SCF types on EF1000

#	SCF	Example	P.	R.	FI	Freq.
68	xcomp_VBARE=>aux_TO	I [prefer] to avoid sitcoms at all.	.93	.98	.96	110
66	xcomp_N	I want to [become] the new president.	.97	.87	.92	192
1	acomp	... helps us [feel] easier.	.86	.95	.90	212
23	dobj_N	I urge you to [consider] it.	.90	.90	.90	643
10	ccomp_VTENSED=>mark-that	Someone [mention] that you are untidy.	.87	.93	.90	28
50	pobj	I [go] to bed at twelve o'clock.	.83	.89	.86	232
42	dobj_N:xcomp_VBARE=>aux_TO	Can I [force] them to fix the house?	.86	.86	.86	35
9	ccomp_VTENSED	I [think] the beige sweater is expensive.	.80	.92	.86	61
69	xcomp_VING	I [like] playing tennis.	.86	.72	.78	25
60	prt	To [sum] up,...	.76	.79	.77	28
41	dobj_N:xcomp_VBARE	[Let] me tell you why ...	.77	.77	.77	13
39	dobj_N:xcomp_ADJ	... to [make] them heavier.	.80	.73	.76	11
24	dobj_N:iobj	Let me [tell] you what I did.	.77	.65	.71	26
65	su	I can drive and [sing].	.68	.74	.71	96
32	dobj_N:pobj	John is going to [tell] Isabella about that.	.70	.68	.69	112
36	dobj_N:prt	... if you [give] up your studies.	.71	.65	.68	26
48	pcomp_VING	... [forget] about asking the prices.	1.0	.44	.62	9
54	pobj:prt	I [look] forward to the start of classes.	.75	.50	.60	12
44	dobj_N:xcomp_VING	I must [spend] four years finishing my university life.	.60	.50	.55	6
34	dobj_N:pobj:prt	... [put] down your ideas as bullet points.	.67	.44	.53	9
18	ccomp_WHCOMP	... you will [learn] how to handle emergent case timely.	.55	.46	.50	13
3	advmod	Then [turn] left at Green Ave.	.35	.62	.44	13
67	xcomp_VBARE	... I [like] take a walk.	.60	.33	.43	9
4	advmod:dobj_N	... [spend] our time there.	.00	.00	.00	6

3.3.1 *Accuracy*

We evaluated the precision, recall and F1 score of individual SCF types during the 10-fold cross-validation of the SCF identification system. The system is able to classify 49 SCF types, which are the union of the SCF types that occurred in both learner data and native data. However, 11 SCF types appeared only in native data (indicated by "n" in the Appendix), which meant we cannot evaluate their accuracy on the learner data. Moreover, some SCFs were rare in the learner data, which made their evaluation unreliable. For example, when a SCF type had only two verb tokens, the training set might include none of the verb tokens, which made it impossible for the model to classify the SCF type correctly. Alternatively, if the training set and testing set had one verb token each, the accuracy scores of this SCF type would be either a hundred or zero per cent, depending on whether the verb token in the testing set was classified correctly or not. Such accuracy rates are uninformative. As a result, we omit 14 SCFs that had fewer than five verb tokens in the learner data (indicated by "r" in the Appendix) from the results. Table 1 lists the remaining 24 SCF types (indicated by "*" in the Appendix), each illustrated with an example from EF1000. The first column denotes the ID number of a SCF in the Appendix, where the guideline examples of the SCFs are available.

As we can see, the majority of the SCF types were classified accurately. Eight SCF types, which accounted for 77% of the learner data, were identified with an F1-score of over 85%. Contrastingly, six SCF types, which accounted for only 3% of the learner data, were identified with an F1-score of less than 60%. To some extent, the low accuracy of the rare SCF types was caused by the scarcity of their training data for the model.

3.3.2 *Error analysis*

We analyzed the identification errors during testing to find out what SCF types were challenging for our system, and to diagnose the cause of the errors. Table 2 lists the SCF misanalysis pairs that occurred at least five times during testing.

The most frequent misanalysis was found between "dobj_N" (a direct object) and "dobj_N:pobj" (a direct object and a prepositional object), which related to the inclusion or exclusion of a prepositional object. Similarly, the misanalysis pair of "su" (intransitive) and "pobj", and the misanalysis between "pobj:pobj" (two prepositional objects) and "pobj", involved a decision about a prepositional object. Further analysis

revealed that there were two main causes of the misidentification of SCFs with regard to prepositional objects.

Table 2. SCF confusion pairs during testing

Target	Prediction	Freq.
dobj_N	dobj_N:pobj	22
dobj_N:pobj	dobj_N	19
xcomp_N	acomp	19
dobj_N	su	12
dobj_N:pobj	pobj	11
su	pobj	10
pobj	dobj_N:pobj	7
dobj_N:iobj	dobj_N	6
dobj_N	ccomp_VTENSED	5
su	dobj_N	5
dobj_N	dobj_N:iobj	5
pobj	su	5
pobj:pobj	pobj	5

i.Distinction between arguments and adjuncts

The SCF identifier erroneously considered the temporal prepositional object in (4) as an adjunct, misidentifying "dobj_N:pobj" as "dobj_N". Even though the verb *do* rarely takes a prepositional object as a complement, and a temporal prepositional object is usually an adjunct, the phrase *in 1874* was a complement of *done* due to the criterion of obligatoriness (Meyers et al., 1996): Example (4) would be ungrammatical if the temporal prepositional object was removed. Example (5) illustrates a SCF misanalysis in the opposite direction: the SCF identifier erroneously included the locational prepositional object *on her birthday party* as a complement, misidentifying "dobj_N" as "dobj_N:pobj".

(4) It's an oil painting [done]$_{*dobj\_N}$ in 1874.

(5) Jane would like to [see]$_{*dobj\_N:pobj}$ you on her birthday party.

(6) The graph [provides]$_{*dobj\_N:pobj}$ sales figures for international sales and ...

(7) What do I wish to [do]$_{*su}$?

Example (4) illustrates the difficulty in distinguishing between complements and adjuncts for prepositional objects. This was caused not only by the limitation of our

model, but also the inherent fuzziness between complements and adjuncts (Somers, 1984).

ii.Prepositional attachment

Another frequent cause of the misidentification errors regarding prepositional objects was prepositional attachment errors. For instance, the prepositional object in (6) should be attached to (i.e. be a dependent of) the noun phrase *sales figure.* However, the SCF identification system erroneously considered the prepositional object as a complement of the predicate *provides,* resulting in the misidentification of "dobj_N" as "dobj_N:pobj". This problem was mainly caused by the errors of the dependency parser. Prepositional attachment is notoriously difficult for NLP.

The SCF identifier also misidentified "xcomp_N" (a nominal complement) as "acomp" (an adjectival complement) sometimes. Further analysis showed that most of such errors happened on nominal complements headed by proper nouns (e.g. *Werner* in *My name is Werner* was misidentified as an adjectival complement). Furthermore, the SCF identifier sometimes omitted a direct object, e.g. misidentifying "dobj_N" as "su" (intransitive) in (7). These errors were mainly found when a direct object preceded the predicate. Such problems were caused by the scarcity of the relevant training cases for the model.

Despite of the aforementioned errors, the SCF identifier is accurate in general and can be useful for linguistic annotation and analysis. We will illustrate this in the next section.

4. Case study: SCF diversity and L2 proficiency

The major advantage of the SCF identification system lies in the scale of SCF data it can produce. The system can facilitate SCF annotation, and can support searching and analyzing SCFs on large-scale corpora. To illustrate the usefulness of the system for learner corpus research and SLA, this section presents an investigation into how L2 learners diversify their use of SCFs in text and how this diversity changes with L2 proficiency.

No research has been conducted to investigate the diversity of SCF use

in L2 learning, as it requires a large amount of SCFs data. Nevertheless, such research has potential value for L2 research and education. First, while it is intuitive to hypothesize that L2 learners can use a wider range of SCFs as their L2 proficiency develops, it is unclear how L2 learners diversify their use of SCFs in text and how this diversity changes across different proficiency levels. For example, do L2 learners repeat fewer SCFs in text when their proficiency improves? How about their distribution of SCFs in text -- do L2 learners distribute different SCFs more evenly as their proficiency improves? Answers to such questions can help researchers to better understand L2 SCF learning, and can assist L2 educators to teach and develop educational material for L2 learners at different stages.

Second, the diversity of SCF use has the potential to contribute to linguistic complexity research. Previous studies have shown that lexical diversity indices are useful predictors of language proficiency (Jarvis, 2013). It is possible that the diversity of lexical and morphosyntaxtic features encoded through SCFs also correlate with language proficiency. Furthermore, researchers are calling for more specific and multidimensional metrics of linguistic complexity, as different dimensions of linguistic complexity may not increase linearly with proficiency and it is more informative to portray them separately (Norris & Ortega, 2009). SCF diversity may contribute a new perspective to the measurement of linguistic complexity[3].

In the rest of this section, we first design multi-dimensional SCF diversity metrics. We then apply the SCF diversity metrics to the learner essays in EFCAMDAT, and investigate the relation between SCF diversity and L2 proficiency.

4.1 Design of SCF diversity metrics

Examples (8) and (9) are extracts from EFCAMDAT. At a first glance, the second extract seems to involve more diverse SCFs than the first one. The question we address in this section is what metrics can be used to reflect such difference in the diversity.

[3] Note that due to the limited space, our investigation of SCF diversity is preliminary. We leave the investigation of the relation between SCF diversity and linguistic complexity to future work.

(8) (Level 4) Hi, Granny. I [feed] $_{dobj\_N}$ the dog at 8am every day and [walk] $_{dobj\_N}$ the dog in the afternoon, after [walk] $_{dobj\_N}$ the dog, I [feed] $_{dobj\_N}$ the dog again at 5pm, every day. Please, [shopping] $_{dobj\_N}$ the dog's food because I [am]$_{acomp}$ tired and I need [do] $_{dobj\_N}$ my homework.

(9) (Level 7) Hello, my name [is] $_{xcomp\_N}$ Saad, I [heard] $_{dobj\_N}$ some rumors about my favorite actor Gavin Taylor and his wife. In my opinion most celebrities are always [exposing] $_{pobj}$ to rumors, especially in their private life. Because all the media and the fans are [following] $_{dobj\_N:prt}$ up their news. Actually, I didnt [see] $_{dobj\_N}$ the TV interview with Taylors wife, but I [think] $_{ccomp\_VTENSED}$ all that news about Gavin Taylor and his wife [are] $_{xcomp\_N}$ rumors from someone to frame Gavin.

4.1.1 *Basic design*

The diversity of a group of elements can be investigated from several dimensions. Inspired by the design of species diversity and lexical diversity metrics (Jarvis, 2013), we designed SCF diversity metrics from four dimensions: repetition, evenness, dispersion and disparity. We chose these dimensions because they were distinct from each other and had intuitive connection with diversity, as we explain below.

i. Repetition

The repetition of SCFs reflects how many SCFs are repeated. The more SCFs are repeated, the less diverse the SCFs are. We measured SCF repetition with the type-token ratio (TTR) of SCFs (hereafter referred to as SCF TTR). For example, both Extract 1 and 2 involve 7 SCF tokens; Extract 1 has only 2 SCF types (SCF TTR=2/7) whereas Extract 2 has 6 (SCF TTR=6/7). This indicates that Extract 2 has a lower SCF repetition.

ii. Evenness

The evenness of SCFs reflects how close in frequency the SCF types are. This dimension of diversity considers the frequency distribution of the SCFs: supposing

the degree of SCF repetition is fixed, the more even that the SCFs are distributed across the types, the more diverse the SCFs are. For example, using simple SCF types frequently is considered less diverse than using simple and complicated SCFs types equally. We measured SCF evenness with the standard deviation of the SCF tokens for each SCF type (hereafter referred to as SD-based SCF evenness):

$$SD = \sqrt{\frac{\sum_{i=1}^{S}\left(n_i - \bar{n}\right)}{S-1}} \quad \text{(b)}$$

where n_i is the number of SCF tokens for the i-th SCF type, and $\bar{n}$ is the average number of SCF tokens across all SCF types. Note that SD requires the presence of two or more SCF types, otherwise the denominator in the formula becomes zero. The lower the standard deviation is, the more evenly that the SCFs are allocated across different SCF types. For example, the SDs of Extract 1 and 2 are 3.53 and 0.41 respectively, indicating that the SCF distribution of Extract 2 is more even.

iii. Dispersion

The dispersion of SCFs reflects how far away the SCF tokens of the same SCF type are located. This dimension of the diversity considers the textual location of the SCFs: supposing the degree of SCF repetition and evenness is fixed, the further away the SCFs of the same types are located from each other, the more dispersed the SCFs are, presenting a higher surface diversity. We calculated SCF dispersion by the average distance between the SCF tokens of the same SCF type:

$$D = \frac{\sum_{j}^{S}\left(\sum_{i=1}^{R_j-1}\left|position_{i,j} - position_{i+1,j}\right| / R_j\right)}{S} \quad \text{(c)}$$

where $position_{i,j}$ refers to the position of the i-th SCF token of the j-th SCF type which has R_j (R_j >= 2) SCF tokens in total. Note that this formula requires the presence of at least two SCFs for a SCF type. We used two kinds of position: the word position in text, and the verb position relative to all verbs (hereafter referred to as word-based SCF dispersion and verb-based SCF dispersion respectively). To illustrate, the word distance between $[\text{feed}]_{\text{dobj\_N}}$ and $[\text{walk}]_{\text{dobj\_N}}$ in Extract 1 is 5 (we calculate punctuation as a word), while the verb distance is 1.

iv. Disparity

The disparity of SCFs reflects how taxonomically different the SCFs are. Some SCF types are more similar than the others. For example, ccomp_VBARE is similar to ccomp_VTENSED because they both have a clausal complement, whereas dobj_N is more different. Supposing the degree of SCF repetition, evenness and dispersion is fixed, the greater the taxonomic distance between the SCFs is, the more diverse the SCFs are.

To measure SCF disparity, we classified the complements by dependency relations, and further classified 4 subtypes for "ccomp" ("VBARE", "VTENSED", "VTENSED=>mark-that", "WHCOMP"), 3 subtypes for "prep" ("pobj", "pcomp", "pcomp=>VING") and 7 subtypes for "xcomp" ("N", "ADJ", "VBARE", "VBARE=>aux_TO", "VEN", "VING", "WHCOMP"). We then calculated the taxonomic distance between two SCF types as follows. First, if the dependency relations of the complements were different, the number of different dependency relations was added to the distance score. Second, for the complements of the same dependency relation, if they had different subtypes, the number of different subtypes was weighted by 0.25 and added to the distance score. For example, the distance between "dobj_N" and "ccomp_VTENSED" is 2, whereas the distance between "ccomp_VTENSED=>mark-that" and "ccomp_VBARE" is 0.5.

We calculated two SCF disparity metrics based on the taxonomic distance: the maximum and average of the pairwise taxonomic distance between SCF types (hereafter referred to as max-based SCF dispersion and AVG-based SCF dispersion respectively). For example, the max-based disparity is 2 for Extract 1 (due to the distance between "dobj_N" and "acomp") and 3 for the Extract 2 (due to the distance between "dobj_N:prt" and e.g. "xcomp_N").

4.1.2 *Control for text length*

The SCF diversity metrics are susceptible to text length. For example, as the text becomes longer, SCF TTR tends to decrease, because the number of SCF tokens increases whereas the increase of SCF types slows down and stops when the writer have used all the types he or she knows. To compare the SCF diversity of texts with different length, we need to control the SCF diversity metrics for text length.

We standardized a SCF metric by calculating the average of the metric for a window of a fixed number of verbs moving across a text. For example, if we set the window size to be five verbs, the first window step for Extract 1 spans from the predicate *feed* to the predicate *shopping*. The window then moves by one verb, with the second step spanning from the predicate *walk* to the predicate *am*. The window moves until it reaches the last predicate in the text, and the scores of all window steps are averaged. This standardization method was inspired by the calculation of mean moving-average type-token ratio (MATTR) for words (Covington & McFall, 2010).

We standardized all SCF diversity metrics over the window sizes of 5, 10 and 20 verbs respectively. When a text had fewer verbs than the window size, the standardized metrics were considered as inapplicable for the text (i.e. the text was excluded from analysis). Obviously, a larger window size applies to fewer texts. We avoided the window size of more than 20 verbs because the number of texts for such window size was small.

The window size influences the properties of the standardized metrics. First, the window size corresponds to the size of linguistic unit for consideration. A smaller window size is closer to the sentence level, whereas a larger window size is related to a larger discourse. Second, a smaller window size makes it easier for the metrics to "saturate", i.e. reach the maximum possible value. For example, it is easier to find completely different SCF types for 5 verbs than 10 verbs. Third, a larger window size leads to a finer granularity for the metrics. For example, the SCF TTR for a window step of 5 verbs can take the value of 0.2, 0.4, 0.6, 0.8, or 1, corresponding to 1 to 5 SCF types within the window, whereas for 10 verbs the value can be 0.1, 0.2, 0.3…., and 1.

For a metric that has a requirement over SCF distribution, e.g. SCF evenness, which requires the presence of two or more SCF types, we considered only the window steps which met the requirement; if no window step met the requirement, the metric was considered as inapplicable for the text.

4.2 Data selection and statistical analysis method

We applied our SCF identifier to the whole EFCAMDAT[4], and calculated the standardized SCF metrics for each text. The L2 proficiency was operationalized as the numerical value of the proficiency levels in EFCAMDAT (i.e. 1-16). To facilitate comparison between different dimensions of SCF diversity, we selected the texts on which all SCF metrics standardized at a particular window size were applicable (e.g. for the window size of 10, we selected texts containing at least 2 SCF types and 2 SCF instances of the same SCF type in a window of 10 verbs so that SCF evenness and dispersion metrics standardized at this window size were applicable), resulting in 508,192, 301,255 and 51,719 texts for the window sizes of 5, 10 and 20 verbs respectively. The three text groups are hereafter referred to as DAT5, DAT10 and DAT20. The number of texts decreased with the window size because, for example, a text containing 6 verbs may be included in DAT 5, but was excluded from DAT10 and DAT20 which required at least 10 and 20 verbs per text respectively. Nevertheless, even the smallest dataset (DAT20) had more than 323 texts for each L2 proficiency level. The size of each dataset was large enough for the statistical analysis reported below. Meanwhile, the texts were distributed unevenly across L2 proficiency levels. As each L2 proficiency level corresponded to 8 writing tasks, we weighed each data point by the inverse of the frequency of the writing task during the statistical analyses to achieve a balanced contribution of residuals across different proficiency level.

We then investigated whether a significantly positive or negative linear relation exists between the SCF diversity metrics and L2 proficiency. While the relation between SCF diversity and L2 proficiency level may be non-linear and might be susceptible to other factors, e.g. topics, writing tasks and L1 influence, we aimed to find a general relation first, which can serve as a starting point for wider analysis. We first checked the scatter plots and line graphs between the SCF diversity metrics and L2 proficiency. As some linear relations were identified, we analyzed the correlation between the SCF diversity metrics and L2 proficiency. We also performed multiple regression analyses (MRA) to investigate how much the combined SCF metrics accounted for the variance in the L2 proficiency level. We selected the SCFs metrics for the MRA as follows: first, we ensured a linear relation between each explanatory variable and the explained variable by choosing the SCF metrics that showed an

[4] As mentioned in Section 4.1, the SCF identifier does not analyze the new frames created by learners. Since the occurrence of such frames was rare (0.6%), this case study assumed that the negligence of creative frames did not affect the result. We encourage future research to investigate the use of creative SCFs, and such data may be achieved by manually editing the output of the SCF identifier.

absolute correlation of $|r| > 0.1$ (the threshold for a small effect, Cohen 1988) with L2 proficiency. Second, we prevented multicollinearity between the explanatory variables by conducting a pairwise correlation test on the selected SCF metrics, and for each pair of SCF metrics that had an absolute correlation of $|r| > 0.7$, we kept the metric that had the highest absolute correlation with L2 proficiency.

4.3 Results

Figure 2 shows how the mean (and its 95% confidence interval) of each SCF metric standardized at the window size of 5 verbs changed with L2 proficiency on DAT5. The figures for the SCF diversity metrics standardized at other window sizes and other datasets are similar. As we can see, there was a near-linear relation between each SCF diversity metric and L2 proficiency.

INSERT FIGURE 2 HERE

Figure 2. Relation between the average of SCF diversity metrics and L2 proficiency (DAT5)

We then analyzed the Pearson correlation between each SCF diversity metric and L2 proficiency. Table 3 shows the results. All correlations were significant at the level of $p < 0.001$, with an achieved power of > 0.999 (i.e., $\beta < 0.001$). This means that the correlations were statically robust. Note that SCF TTR and disparity metrics standardized at a smaller window size are applicable to the datasets prepared for a larger window size, but this is not necessary the case for SCF evenness and dispersion metrics, which have requirement on SCF distribution (e.g. SCF dispersion requires the presence of at least two SCF types in the standardization window): for a text where such requirements are satisfied at a large window, the requirement may not be satisfied at a smaller window. In our case, the SCF dispersion metrics standardized at the window size of 5 verbs were inapplicable to DAT10.

As we can see from Table 3, SCF TTR showed medium positive correlations ($r > 0.3$) with L2 proficiency. This meant that more advanced learners tended to repeat fewer SCFs, and this trend can be observed at all window sizes that we investigated. Meanwhile, the correlation increased with the standardization window size on DAT10 and DAT20. For example, on DAT20, SCF TTR standardized at the

window size of 5 verbs showed a correlation of 0.315, whereas for 20 verbs the correlation was 0.355. This might be attributed to the following factors: Frist, SCF TTR for a larger linguistic unit might reflect the increase in L2 proficiency better. Second, the metric standardized at a larger window size had a lower rate of saturation and finer granularity, which made the metric more informative.

Table 3. Correlation between standardized SCF metrics (repetition and disparity) and L2 proficiency

Metrics	**Window size (# of verbs)**	**Dataset**		
		DAT5	**DAT10**	**DAT20**
SCF TTR	5	.368	.353	.315
	10	--	.357	.352
	20	--	--	.355
SD-based SCF evenness	5	-.306	-.292	-.256
	10	--	-.240	-.235
	20	--	--	-.223
Word-based SCF dispersion	5	.248	--	.331
	10	--	.314	.310
	20	--	--	.290
Verb-based SCF dispersion	5	.183	--	.139
	10	--	.190	.153
	20	--	--	.159
Max-based SCF disparity	5	.364	.389	.369
	10	--	.376	.380
	20	--	--	.281
AVG-based SCF disparity	5	.275	.288	.267
	10	--	.257	.272
	20	--	--	.238

SCF evenness had close-to-medium positive correlations ($0.1 < r < 0.3$) with L2 proficiency (note that a lower SCF SD value means higher SCF evenness), except when standardized at the window size of 5 verbs, which reached a correlation of 0.307. This meant that more advanced learners used different SCF types more evenly. Moreover, the correlation on DAT10 and DAT20 was higher when SD-based SCF evenness was standardized at a smaller window size. This meant that unlike SCF TTR, SCF SD reflects the increase in L2 proficiency better at a smaller window size.

As SCF dispersion, word-based SCF dispersion showed medium positive correlations with L2 proficiency, whereas verb-based SCF dispersion showed small positive correlations. This meant that more advanced learners located the verb tokens of the same SCF type further away from each other, and the effect was more obvious when the distance was evaluated by words rather than verbs. This meant advanced learners used more words between the verbs, a result in line with the previous

findings that the mean length of utterance increases with proficiency, and that more advanced learners use more modifiers in noun phrases (Biber et al. 2011; Kyle 2016; Taguchi et al. 2013).

SCF disparity showed medium or close-to-medium positive correlations with L2 proficiency. This meant that more advanced learners used more taxonomically different SCFs. Meanwhile, max-based SCF disparity showed a stronger correlation with L2 proficiency than AVG-based SCF diversity. The former reached 0.389 when standardized at the window size of 5 verbs and applied to DAT10. However, the correlation dropped by almost 0.1 when max-based SCF diversity was standardized at the window size of 20 verbs rather than 10 verbs. This meant that the maximum taxonomic difference between SCFs across a larger linguistic unit is less indicative of L2 proficiency. This is probably because the chance of having taxonomically different SCF types is higher in a larger linguistic unit, and the maximum taxonomic difference becomes similar across different proficiency levels.

Finally, our MRA revealed that the SCF diversity metrics can explain 18.8%, 19.8% and 25.1% of the variance in L2 proficiency on DAT5, DAT10 and DAT20 respectively. Table 4 shows the coefficients of the predictor variables of each model. The increasing explanatory power on the datasets prepared for a larger window size is attributed to the fact that some SCF metrics standardized at a smaller window size were also selected, increasing the number of explanatory variables. Meanwhile, SCF TTR, max-based SCF disparity and word-based SCF dispersion were selected by all models. These metrics represented the unique aspects of SCF diversity that best predicted L2 proficiency[5].

Table 4. Coefficients of the predictor variables in the MRA

Dataset	**Predictor variable (window size)**	**Std. Coefficients**	**t**	***p***
DAT5	SCF TTR (5)	.264	98.550	< .001
	Word-based SCF dispersion (5)	.217	93.507	< .001
	Max-based SCF disparity (5)	.176	72.328	< .001
	Verb-based SCF dispersion (5)	-.150	-57.996	< .001
DAT10	Max-based SCF disparity (5)	.207	84.718	< .001
	Word-based SCF dispersion (10)	.186	97.116	< .001

[5] EFCAMDAT has data from adult learners, some of whom could be false beginners. As one anonymous reviewer points out, it would be interesting to investigate whether the same result holds on the corpora which primarily include absolute beginners (e.g. ICCI or JEFLL) in future research.

	SCF TTR (10)	.167	70.602	< .001
DAT20	Word-based SCF dispersion (5)	.331	128.001	< .001
	SCF TTR (20)	.248	102.666	< .001
	Max-based SCF disparity (10)	.189	83.785	< .001
	Verb-based SCF dispersion (5)	-.155	-63.674	< .001
	Word-based SCF dispersion (20)	.041	19.483	< .001
	AVG-based SCF disparity (10)	-.032	-13.229	< .001

5. Conclusion

We presented the first SCF identification system for learner English, which can label the SCFs of individual verb tokens in text for a set of 49 distinct SCFs at an accuracy of 84.2%. This level of accuracy was among the highest reported of contemporary systems and was likely to be sufficient for benefit in downstream tasks.

The system can support SCF annotation and L2 SCF research based on large-scale corpora. To illustrate, we proposed the first multidimensional SCF diversity metrics and investigated how SCF diversity changed with L2 development. Our results shed light on L2 SCF acquisition: more advanced learners tended to use more diverse SCF types which were taxonomically more different from each other; meanwhile, more advanced learners tended to use different SCF types more evenly, and locate the verb tokens of the same SCF type further away from each other.

Our SCF identification system opens up many opportunities for linguistic research. For example, researchers can investigate how SCF use changes across different L1 backgrounds, and whether there is any L1 transfer on L2 SCF use from the typological aspect. Furthermore, task effects are widely recognized as an important aspect of learner language analysis (Alexopoulou et al., 2017), and it will be interesting to investigate how writing tasks affect SCF use. Moreover, researchers can include SCFs or SCF diversity into the design of course materials and language assessment.

References

Al-Rfou', R., Perozzi, B., & Skiena, S. (2013). Polyglot: Distributed word representations for multilingual NLP. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 183–192. https://www.aclweb.org/anthology/W13-3520

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, *67*(S1), 180–208.

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., & Collins, M. (2016). Globally normalized transition-based neural networks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2442–2452. https://doi.org/10.18653/v1/P16-1231

Aston, G., & Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.

Baker, S., Reichart, R., & Korhonen, A. (2014). An unsupervised model for instance level subcategorization acquisition. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 278–289.

Berger, A. L., Pietra, V. J. Della, & Pietra, S. A. Della. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, *22*(1), 39–71.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press.

Boguraev, B., & Briscoe, T. (1987). Large lexicons for natural language processing: Utilising the grammar coding system of LDOCE. *Computational Linguistics*, *13*(3–4), 203–218.

Briscoe, T., & Carroll, J. (1997). Automatic extraction of subcategorization from corpora. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 356–363.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). John Benjamins.

Charniak, E., & Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In K. Knight (Ed.), *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 173–180).

Chen, X., & Meurers, D. (2019). Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning*, *32*(4), 418–447.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Earlbaum Associates.

Council of Europe. (2001). *Common European Framework of Reference for Languages:*

Learning, Teaching, Assessment. Cambridge University Press.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100.

De Marneffe, M.-C., & Manning, C. D. (2008). The Stanford typed dependencies representation. *COLING 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, 1–8.

Dušek, O., Hajič, J., & Urešová, Z. (2014). Verbal valency frame detection and selection in Czech and English. *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, 6–11. https://doi.org/10.3115/v1/W14-2902

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based Approaches to Language Acquisition and Processing: Cognitive and Corpus Investigations of Construction Grammar*. Wiley.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In R. T. Miller, K. I. Martin, C. M. Eddington, A. Henery, N. M. Miguel, A. Tseng, A. Tuninetti, & D. Walter (Eds.), *Proceedings of the 31st Second Language Research Forum: Building Bridges Between Disciplines*. Cascadilla Proceedings Project.

Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). SimVerb-3500: A large-Scale evaluation set of verb similarity. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2173–2182. https://aclweb.org/anthology/D16-1235

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234.

Gries, S. T., & Berez, A. L. (2017). Linguistic annotation in/for corpus linguistics. In N. Ide & J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation* (pp. 379–409). Springer.

Grishman, R., Macleod, C., & Meyers, A. (1994). COMLEX syntax: Building a computational lexicon. *Proceedings of the 15th Conference on Computational Linguistics-Volume 1*, 268–272.

Helbig, G., & Schenkel, W. (1991). *Wörterbuch zur Valenz und Distribution deutscher Verben*. VEB Bibliographisches Institut.

Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics*, *23*(1), 28–54.

Jackendoff, R. (n.d.). *Semantic structures*. MIT press.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, *63*(s1), 87–106.

Kyle, K. (2016). *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-based Indices of Syntactic Sophistication* [Doctoral dissertation, Georgia State University]. https://scholarworks.gsu.edu/alesl_diss/35/

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Meurers, D., Krivanek, J., & Bykh, S. (2013). On the automatic analysis of learner corpora: Native language identification as experimental testbed of language modeling between surface features and linguistic abstraction. In A. A. Sintes & S. V. Hernández (Eds.), *Diachrony and Synchrony in English Corpus Studies*. Peter Lang.

Meyers, A., Macleod, C., & Grishman, R. (1996). Standardization of the complement adjunct distinction. *Proceedings of EURALEX 96 (International Conference on Lexicography)*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (Eds.), *1st International Conference on Learning Representations*. http://arxiv.org/abs/1301.3781

Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In A. Dawn, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 conference* (pp. 572–581). UCREL.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578.

Preiss, J., Briscoe, T., & Korhonen, A. (2007). A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 912–919.

Procter, P. (1978). *Longman dictionary of contemporary English*. Longman.

Quochi, V., Frontini, F., Bartolini, R., Hamon, O., Poch, M., Padró, M., Bel, N., Thurmair, G., Toral, A., & Kamram, A. (2014). *Third Evaluation Report. Evaluation of PANACEA v3 and Produced Resources*. http://hdl.handle.net/10230/22533

Römer, U., O'Donnell, M. B., & Ellis, N. C. (2015). Using COBUILD grammar patterns for a large-scale analysis of verb-argument constructions. In *Corpora, grammar and discourse: In honour of Susan Hunston* (Vol. 73, pp. 43–72). John Benjamins.

Römer, U., Roberson, A., O'Donnell, M. B., & Ellis, N. C. (2014). Linking learner corpus

and experimental data in studying second language learners' knowledge of verb-argument constructions. *ICAME Journal*, *38*(1), 115–135.

Somers, H. L. (1984). On the validity of the complement-adjunct distinction in valency grammar. *Linguistics*, *22*(4), 507–530.

Tesnière, L. (1965). *Eléments de Syntaxe Structurale*. John Benjamins.

Tono, Y. (2004). Multiple comparisons of IL, L1 and TL corpora: The case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. *Corpora and Language Learners*, *17*, 45–66.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. University of Hawai'i Press.

Appendix. SCF inventory and examples

		SCF	Examples
1	*	acomp	His reputation sank low. He appears crazy / distressed. He seems well.
2		acomp:prt	He started out poor.
3	*	advmod	He meant well. It carves easily.
4	*	advmod:dobj_N	He put it there. They mistakenly thought him here.
5		advmod:prt	He came off badly.
6	n	ccomp_VBARE=>mark-that	She demanded that he leave.
7		ccomp_VBARE=>mark-that:iobj	He petitioned them that he be freed.
8		ccomp_VBARE=>mark-that:pobj	They suggested to him that he go.
9	*	ccomp_VTENSED	They thought he was always late. He seems as if he is clever.
10	*	ccomp_VTENSED=>mark-that	To report the theft indicates that he wasn't guilty. It seems that they left. He complained that they were coming.
11	r	ccomp_VTENSED=>mark-that:dobj_N	It annoys them that she left. I take it that Kim left. It is believed that he came.
12		ccomp_VTENSED=>mark-that:dobj_N:iobj	He bet her ten pounds that he came
13		ccomp_VTENSED=>mark-that:dobj_N:prt	He had her on that he attended.
14	n	ccomp_VTENSED=>mark-that:iobj	He told the audience that he was leaving.
15	r	ccomp_VTENSED=>mark-that:pobj	It matters to them that she left. They admitted to the

			authorities that they had entered illegally.
16		ccomp_VTENSED=>mark-that:pobj:prt	She gets through to him that he came.
17	n	ccomp_VTENSED=>mark-that:prt	They figured out that she hadn't done her job. It turns out that he did it.
18	*	ccomp_WHCOMP	He asked how she did it. He asked whether he should come. He asked what he should do.
19	r	ccomp_WHCOMP:dobj_N	I would appreciate it if he came.
20	r	ccomp_WHCOMP:iobj	They asked him whether he was going. They asked him what he was doing. He asked him how he came.
21		ccomp_WHCOMP:pobj	He explained to her how she did it. They asked about everybody whether they had enrolled. They asked about everybody what they had done. It dawned on him what he should do.
22	n	ccomp_WHCOMP:prt	They figured out whether she hadn't done her job. They figured out what she hadn't done.
23	*	dobj_N	That she left annoys them. To read pleases them. He loved her. He combed the woods looking for her. It cost ten pounds.
24	*	dobj_N:iobj	She asked him his name. It cost him ten pounds.
25		dobj_N:iobj:pobj:xcomp_VBARE=>aux_TO	It cost Kim a pound for us to go.
26		dobj_N:iobj:prt	I opened him up a new bank account It set him back ten pounds
27		dobj_N:iobj:xcomp_VBARE=>aux_TO	It took us an hour to find.
28	r	dobj_N:pcomp	They helped me with whatever I was doing. He strikes me as foolish. He condemned him as stupid. He accepted him as associated. He accepted him as being normal.
29		dobj_N:pcomp:prt	He put him down as stupid
30	n	dobj_N:pcomp_VING	I prevented her from leaving. I accused her of murdering her

			husband. He wasted time on fussing with his hair. He told her about climbing the mountain. They asked him about his participating in the conference. He attributed his failure to no one buying his books.
31		dobj_N:pcomp_VING:prt	He talked him around into leaving
32	*	dobj_N:pobj	I sent him as a messenger. She served the firm as a researcher. She bought a book for him. She added the flowers to the bouquet. I considered that problem of little concern. He gave a big kiss to his mother. He made use of the money.
33	r	dobj_N:pobj:pobj	He turned it from a disaster into a victory
34	*	dobj_N:pobj:prt	I separated out the three boys from the crowd.
35	n	dobj_N:pobj:xcomp_VBARE=>aux_TO	I arranged it with Kim to meet. It requires ten pounds for him to go.
36	*	dobj_N:prt	I looked up the entry.
37		dobj_N:prt:xcomp_ADJ	He makes him out crazy. He sands it down smooth.
38	n	dobj_N:prt:xcomp_VBARE=>aux_TO	He made him out to be crazy. He spurred him on to try.
39	*	dobj_N:xcomp_ADJ	He painted the car black. She considered him foolish.
40	r	dobj_N:xcomp_N	They appointed him professor.
41	*	dobj_N:xcomp_VBARE	He made her sing. He helped her bake the cake.
42	*	dobj_N:xcomp_VBARE=>aux_TO	It pleases them to find a cure. I advised mary to go. John promised mary to resign. They badgered him to go. I found him to be a good doctor.
43	r	dobj_N:xcomp_VEN	He wanted the children found
44	*	dobj_N:xcomp_VING	I kept them laughing. I caught him stealing.
45		iobj:xcomp_WHCOMP	He asked him whether to clean the house. He asked him what to do.
46	r	pcomp	He thought about whether he wanted to go. He thought about what he

			wanted. He thought about whether to go. He thought about what to do.
47		pcomp:pobj	I agreed with him about whether he should kill the peasants. I agreed with him about what he should do. I agreed with him about what to do. I agreed with him about whether to go.
48	*	pcomp_VING	They failed in attempting the climb. They disapproved of attempting the climb. They argued about his coming.
49	n	pcomp_VING:prt	He got around to leaving.
50	*	pobj	I worked as an apprentice cook. That she left matters to them. They worried about him drinking. They apologized to him. The matter seems in dispute.
51	r	pobj:pobj	They flew from London to Rome.
52		pobj:pobj:prt	He came down on him for his bad behavior.
53		pobj:pobj:xcomp_VBARE=>aux_TO	They contracted with him for the man to go.
54	*	pobj:prt	She looked in on her friend.
55		pobj:prt:xcomp_VBARE=>aux_TO	He kept on at him to join.
56	r	pobj:xcomp_VBARE	He looked at him leave
57	r	pobj:xcomp_VBARE=>aux_TO	It remains for us to find a cure. It occurred to them to watch. I prefer for her to do it. He conspired with them to do it. He beckoned to him to come. She appealed to him to go. He appeared to her to be ill.
58	n	pobj:xcomp_VING	She attributed his drinking too much to his anxiety. They limited smoking a pipe to the lounge.
59		pobj:xcomp_WHCOMP	He explained to them how to do it. They deduced from Kim whether to go. They deduced from Kim what to do.

60	*	prt	She gave up.
61	n	prt:xcomp_N	He turned out a fool.
62	r	prt:xcomp_VBARE=>aux_TO	He turned out to be a crook. He set out to win.
63	n	prt:xcomp_VING	He ruled out paying her debts.
64		prt:xcomp_WHCOMP	They figured out whether to go. They figured out what to do.
65	*	su	He went. They met. That he came matters. To see them hurts. It rains.
66	*	xcomp_N	He seemed a fool.
67	*	xcomp_VBARE	He helped bake the cake. He dared dance.
68	*	xcomp_VBARE=>aux_TO	It remains to find a cure. He helped to save the child. He seemed to come. I wanted to come.
69	*	xcomp_VING	His hair needs combing. She stopped smoking. She discussed writing novels. He dismissed their writing novels.
70	r	xcomp_WHCOMP	He explained how to do it. He asked whether to clean the house. He asked what to do.

Address for correspondence

Yan Huang

Language Technology Lab, Faculty of Modern and Medieval Languages and Linguistics

University of Cambridge

Faculty of English Building, 9 West Road

Cambridge, CB3 9DB

United Kingdom

yh358@cantab.ac.uk

Co-author information

Akira Murakami

Department of English Language and Linguistics Institution

University of Birmingham

Birmingham, B15 2TT
United Kingdom
a.murakami@bham.ac.uk

Theodora Alexopoulou
Language Technology Lab, Faculty of Modern and Medieval Languages and Linguistics
University of Cambridge
Faculty of English Building, 9 West Road
Cambridge, CB3 9DB
United Kingdom
ta259@cam.ac.uk

Anna Korhonen
Language Technology Lab, Faculty of Modern and Medieval Languages and Linguistics
University of Cambridge
Faculty of English Building, 9 West Road
Cambridge, CB3 9DB
United Kingdom
alk23@cam.ac.uk