Extracting and Analyzing English Multi-word Expressions with Slots: A Case Study of '*take*'

Tomoyuki Tsuchiya
Kyushu University
tsuchiya@flc.kyushu-u.ac.jp

1 Introduction

Multi-word expression (MWE) has been argued as one of the significant issues for large-scale natural language processing [1], and similar issues have also been coming into discussions in corpus linguistics and cognitive linguistics areas. Conventionalized expressions have been investigated from various aspects– syntactically, semantically and sociolinguistically for instance– and many terms have been applied to this linguistic phenomenon (ex. idioms, formulaic language, construction). One of these terms called 'construction' refers to a type of MWEs whose parts are not fully fixed. This paper discusses the theoretical issues of construction and reports the results of a preliminary investigation using a large-scale web corpus.

Construction Grammar (CxG) is a highly explanatory linguistic theory that aims to reveal the knowledge and productivity of human language based on the speakers' actual usage. CxG assumes conventionalized sequences of linguistic elements called **construction** and one or more empty **slots** in each construction to accept certain lexical items to fit in and generate speech with a certain level of productivity. The linguistic unit of construction can vary into several levels; word, morpheme, or other smaller units, and the level of abstraction of a slot can also vary; part-of-speech, semantics, morphology, and grammatical level. Moreover, many constructions are conventionalized and shared in the language community with nonliteral meaning (ex. *let alone* in "Fred will not eat shrimp, *let alone* squid" from [2]).

This paper investigates a method to extract constructions with slots with certain selection restrictions and word range that are linguistically objective from large-scale corpus data.

1.1 Constructions in Linguistics: A Frame and slots

CxG aims to explain the productivity of language while maintaining the usage-based aspect of language.

The collostructional analysis established by [3] uses a statistical method to investigate each lexical item's productivity that fits in the slot. The term 'collostruction' designates relatively more concrete expressions closer to collocation.

1.2 Remaining Issues of CxG and collostructional analysis

1.2.1 Frame Range of Constructions

Methodologically, collostructional analysis focuses mainly on the type and collocational strength of collexemes but the definition of the starting and end points of each collocation is arbitrary. For instance, [3] conducted a case study of [*X* think nothing of V_{gerund}] and calculated the collocational strength of each collexeme. The word sequence of this study starts from the subject noun phrase as X and ends with gerund verb. [3] investigated this sequence since it appears in the dictionary and V_{gerund} slot has productivity to a certain extent, so designating X as the starting point and V slot as the end point of the sequence should be taken into consideration.

1.2.2 Slots of Constructions

Moreover, the collocational strength indicates the productivity of the slots of a collostruction, however, the slots of the construction are also designated arbitrarily by the researchers themselves. For instance, [*X* think nothing of V_{gerund}] has two empty slots; X and V_{gerund}, but it is more or less possible that the word *nothing* can become a slot to form a more abstract construction [*X* think *Y* of V_{gerund}].

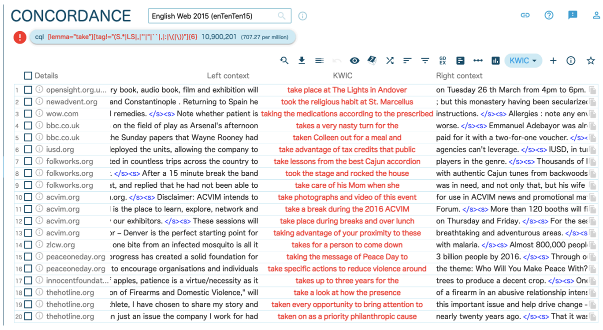

Figure 1 The concordance list of sequences start with *take*

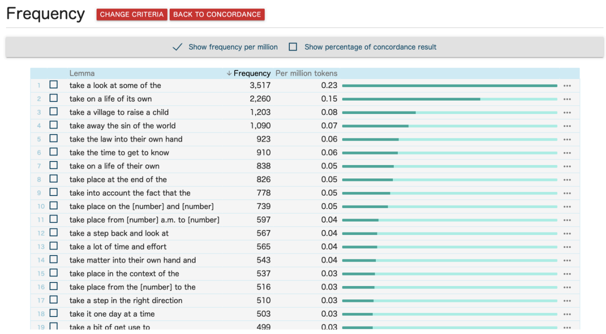

Figure 2 The frequency list of tokens from the concordance

1.2.3 Towards objective designation of the frame range and the slots

By establishing an objective and simple method to extract constructions with productive slots, it would be possible to offer an index that contains useful constructions for language learners.

2 Data and Method

2.1 Data: EnTenTen Corpus

[1] argue the lexical proliferation problem caused by light verb construction with examples; *take a walk, take a hike, take a trip,* and *take a flight*. In order to assess the plausibility of light verb constructions as MWE, this paper focuses on MWEs which starts from a light verb *take*.

For collecting data, the EnTenTen15 corpus was used in this study. EnTenTen15 corpus is an English web corpus, which collects text from the Internet. It contains 13 billion words, and the part-of-speech was tagged on each word by TreeTagger (version 3) using Penn Treebank tagset.

2.2 Method

Using Corpus Query Language (CQL) in the Sketch Engine, expressions consist of 7 words and start from the light verb *take* was collected from the corpus. Any punctuations and symbols are not included in this word sequence. The first 10 million hits were used for the observation due to the Sketch Engine's system default. In most cases, this number is sufficiently representative, so random sampling was not applied in this data collection process.

Subsequently, by counting the frequency of tokens with the Sketch Engine's frequency function, 8,950,602 tokens were found in total. In the counting process, any inflection of words (e.g. declension of nouns and conjugation of verbs) was ignored. From this frequency data, 57 patterns that have at least one slot were made for each token as table 1. Two or more adjacent slots were combined into one slot. For each pattern, type frequency (`t_freq`), total token frequencies (`freq_sum`), the mean of type frequencies (`mean_freq`), and the percentage of the most frequent token in the total token frequency (`max_freq_perc`) were calculated.

Table 1 Slot generation patterns

pattern	verb	R1	R2	R3	R4	R5	R6
base	take	a	step	in	the	right	direction
pattern01	take	...	step	in	the	right	direction
pattern02	take	a	...	in	the	right	direction
pattern02	take	...	...	in	the	right	direction
...	...	...	...	...	...	...	...
pattern56	take	...	...	...	...	right	
pattern57	take	...	...	...	...	...	direction

Finally, patterns that match the following three conditions were extracted from the pattern data; more than three types, more than 1,000 total token frequency, and the most frequent token accounts for more than 10 percent of the total token frequency. Though the constant of each condition above is provisional, it reflects the common characteristic of constructions. The first condition was set in order to extract constructions that have appropriate level of productivity. The second condition was set because constructions with a certain productivity should occur in a certain amount in language activities. The third condition was set since collocational analyses by [3] indicate the existence of a prototypical lexical item, which commonly fits in the slot.

Regarding the procedue above, three types of data have been generated; the raw data which was collected from the actual corpus, the frequency tokens of the raw data, and the abstract patterns that generated from each token. The examples of each data type are as follows.

1. **Raw data**: . . . *film and exhibition will* ***take place at The Lights in Andover*** *on Tuesday 26. . .*
2. **Frequency token**: *take place at The Lights in An-*

dover, *take place during break and over lunch*, *take time out of his busy schedule*

3. **Abstract pattern**:

   patterns from *take place at The Lights in Andover...*
   - *take ... at ... Andover*,
   - *take ... at ... in*,
   - *take ... at ... in Andover*
   - ...
   - *take ... Andover*

   patterns from *take time out of his busy schedule...*
   - *take time ... of his busy schedule*
   - *take time out ... his busy schedule*
   - *take time out of ... busy schedule*
   - ...
   - *take ... schedule*

3 Results and their observation

Consequently, there were 1,693 patterns that meet all of the conditions above. The examples of the extracted patterns are listed in table 3.

Table 2 Basic statistics of the extracted patterns

	mean	min	max	median	stddev
t_freq	694.11	4	19,667	506	1,034.88
freq_sum	2,006.68	1,000	30,137	1,464	1,624.84
max_freq	558.46	102	3,517	328	662.36

As shown in table 4, "take a ... of" has the most types among all patterns and appears the most in the corpus. Most patterns have a highly prototypical token, which account for more than 90% of the tokens appear with the pattern (ex. *take a look ... some of the*, *take ... life ... its*), but there are patterns whose prototypical token accounts for less than 90% (ex. *on a life ... own*, *take ... law into ... own hand*). The tokens appear in the pattern *on a life ... own* were *take on a life {of its/their/his/her/all its/on its/on their} own*, and the tokens appear in the pattern *take ... law into ... own hand* were *take {the/immigration/international/federal/religious} law into {their/his/your/its/our/her/one/my/} own hand.*

4 Discussion and Conclusion

As in table 4, the raw frequency of abstract patterns is higher compared to less abstract patterns. Abstract constructions appear in the corpus more frequently, but the selection restriction of their slots is loose, so various lexical items are acceptable. Therefore, the type frequency of actual expression is higher and do not have a prototypical token.

We can also estimate the end point of the construction by analyzing pattenrs that shares a common prototype token but have different endpoints. For example, among the patterns that share the prototypical token *take away the sin of the world,* the patterns which end with "world" have fewer types than ones that end with "the" in table 6. In terms of the conventionality of language, the patterns that end with "world" are more conventionalized and restrict the repertory of lexical items that fit in the remaining slots.

The slot-ness of each pattern can be estimated by the number of types that appear in the token list. In table 5, for instance, "your ... to ... level" has 904 types in the token list, whereas "your ... the ... level" has approximately 200 types fewer than the former pattern. Though it is not statistically tested, we can estimate that the first pattern has more productive slots than the second one.

This paper pointed out the theoretical difficulties of CxG in terms of the arbitrariness of the designation of the word range and slots and investigated an objective method to designate them by analysing the frequencies of word sequences which start from a light verb *take*. To extract constructions that are beneficial for English learning and create an index of it, the data shown in this paper should be observed in terms of their productivity.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP17KT0061 and JP17K17943.

References

[1] Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for nlp. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 1–15. Springer, 2002.

[2] Charles J Fillmore, Paul Kay, and Mary C. O'onnor. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, Vol. 64, No. 3, pp. 501–538, 1988.

[3] Anatol Stefanowitsch and Stefan Th Gries. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, Vol. 8, No. 2, pp. 209–243, 2003.

Table 3 Pattern examples and basic statistics (15 ≤ `t_freq` ≤ 17)

pattern	t_freq	freq_sum	mean_freq	max_freq	max_freq_perc
a look . . . some of the	15	3590	239.33	3517	97.97%
on a life . . . own	14	3180	227.14	2260	71.07%
. . . a . . . life . . . its	15	2291	152.73	2260	98.65%
. . . life of its own	16	2279	142.44	2260	99.17%
. . . law into . . . own hand	17	1411	83.00	923	65.41%
matter into . . . hand and	17	1264	74.35	543	42.96%
matter into . . . own hand and	15	1262	84.13	543	43.03%
. . . sin . . . world	16	1140	71.25	1090	95.61%
away the . . . world	17	1107	65.12	1090	98.46%
away . . . of . . . world	15	1104	73.60	1090	98.73%
time out of . . . busy schedule	16	1052	65.75	394	37.45%
the . . . to get . . . know	14	1004	71.71	910	90.64%
. . . law . . . their own	17	1001	58.88	923	92.21%

Table 4 The list of patterns with the highest `t_freq` frequency (5,000 ≤ `t_freq`)

pattern	t_freq	freq_sum	mean_freq	max_freq	max_freq_perc
a . . . of	19667	30137	1.53	3517	11.67%
. . . its	19101	22347	1.17	2260	10.11%
. . . some	12106	18005	1.49	3517	19.53%
. . . own	10655	15513	1.46	2260	14.57%
. . . child	8672	11357	1.31	1203	10.59%
. . . life	8497	12871	1.51	2260	17.56%
away . . . of	7094	8922	1.26	1090	12.22%
the . . . into	6547	8844	1.35	923	10.44%
away . . . the	5874	7363	1.25	1090	14.80%
. . . of their	5866	7667	1.31	838	10.93%
. . . know	5308	7421	1.40	910	12.26%

Table 5 A group of patterns that share a common prototype

pattern	t_freq	freq_sum	mean_freq	max_freq	max_freq_perc
your . . . level	1072	3744	3.49	436	11.65%
your . . . to . . . level	904	3562	3.94	436	12.24%
your . . . the . . . level	687	3212	4.68	436	13.57%
your . . . to the . . . level	666	3189	4.79	436	13.67%
your . . . next level	658	3181	4.83	436	13.71%
your . . . to . . . next level	629	3150	5.01	436	13.84%
your . . . the next level	627	3148	5.02	436	13.85%
your . . . to the next level	613	3132	5.11	436	13.92%

Table 6 A group of patterns that share *take away the sin of the world* as the prototype (`t_freq` ≥ 30)

pattern	t_freq	freq_sum	mean_freq	max_freq_perc
away . . . sin of . . . world	4	1093	273.25	99.73%
away . . . sin . . . the world	5	1094	218.80	99.63%
away . . . sin . . . world	6	1095	182.50	99.54%
. . . the sin of the world	7	1131	161.57	96.37%
. . . the sin . . . the world	9	1133	125.89	96.20%
. . . the sin of . . . world	10	1134	113.40	96.12%
. . . sin of the world	10	1134	113.40	96.12%
. . . the sin . . . world	12	1136	94.67	95.95%
. . . sin of . . . world	13	1137	87.46	95.87%
. . . sin . . . the world	13	1137	87.46	95.87%
. . . sin . . . world	16	1140	71.25	95.61%
away . . . sin of the	28	1144	40.86	95.28%
away the sin . . . the	29	1145	39.48	95.20%