How can we find syntactic patterns?:

A case study of English PDC as an embodiment of usage-based view of language[*†]

Masato YOSHIKAWA

1. Introduction

In English, we have three ways of expressing "the transfer of objects (and other things) between people" (Campbell & Tomasello 2000: 253). They are 1) *to*-dative Construction, 2) *for*-dative Construction (or, Benefactive Construction), and 3) Ditransitive (or, Double object) Construction. The former two are often called Prepositional Dative Construction (PDC, hereinafter). Examples of these constructions are as follows:

(1) a. John gave the book to me. (to-dative Construction)
b. John bought the book for me. (for-dative Construction)
c. John gave me the book. (Ditransitive Construction)

We can easily state that these constructions actually exist since we do recognize a certain array of words or phrases, e.g., (1)a, as a single unit, e.g., *to*-dative Construction, not as a mere sequence of items. However, there arises one problem: how can we recognize the constructions in the course of language processing? This is crucial in that unless we solve the problem, we cannot explain the mechanism of language acquisition and processing thoroughly.

---

[*] I would like to express my gratitude to the following people because their comments on the presentation about this topic and discussions with them deeply contributed to the elaboration of the quality of this paper: Professor Yukio Tsuji (Keio University), Professor Ippei Inoue (Keio University), Noburo Saji (Keio University), Fuminori Nakamura (Keio University), Ken-ya Nishikawa (Keio University), Tomohiro Sakai (JSPS Research Fellowship for Young Scientist), and Hiroi Kubota (Keio University). In addition, I would also like to acknowledge Professor William Snell (Keio University) for his proofreading of the original draft of this paper. However, of course, all the remaining deficiencies are, if any, due to my lack of ability.

[†] The data discussed in this paper are the same as those in the poster presentation at the 9th conference of Japanese Cognitive Linguistics Association held at Nagoya University on September 14, 2008, titled "Koubun wa naze chikaku kanou ka? [Why can we recognize constructions?]". In addition, the paper which is to appear in the proceeding of the conference (to be published in 2009) will share some of the theoretical claim with this paper.

Therefore, in this paper, we will address the issue of finding syntactic patterns and provide a possible answer to it. In order to attain this goal, a statistical analysis was performed under a hypothesis that strong co-occurrences within a certain context have great impacts on human cognition; that is, when two or more elements co-occur in a certain context frequently enough, we can recognize the context as a single unit.

2. About Prepositional Dative Constructions (PDC)

2.1. Definition

In the field of Construction Grammar (CG: e.g., Goldberg 1995) and Radical Construction Grammar (RCG: e.g., Croft 2002), PDC has been characterized as follows:

(2) Subj Verb Obj $\text{Obl}_{\text{to/for}}$ (c.f., Goldberg 1995)

(3) a. DA-Subj DA-Verb DA-Obj *to* DA-Loc

b. DA-Subj DA-Verb DA-Obj *for* DA-Benef (c.f., Tomasello 2003)

(2) is a CG's way of description and (3) is that of RCG. The notations *Subj*, *Verb*, *Obj*, *Obl*, *Loc*, and *Benef* represent subject, verb, object, oblique, location, and benefactive, respectively. In (3), the prefix "DA-," attached to all the items but the concrete word *to*, means that these items are unique to Dative Construction and are different from any elements belonging to other constructions such as Transitive Construction, Intransitive Construction, Locative Construction, and so on. Here we adopt the latter way of characterization as a structural definition of PDC.

2.2. Problems

Only from the definition, however, we cannot determine whether a certain array of items is PDC or not; we need further information which tells us that items within a confronting sequence of words or phrases form a larger unit as a whole.

PDC consists of the categorical pattern NP-V-NP-*to/for*-NP, which embodies (3), but we cannot identify as PDC all the sequences of NP-V-NP-*to/for*-NP. For example, such a sentence as follows is never considered as PDC:

(4) John broke the atmosphere to some extent.

This is not regarded as PDC probably because the phrase "to some extent" seems independent of the verb phrase "broke the atmosphere"; that is, it doesn't seem to be a single unit.

In addition, we also have to specify which verb is *DA-Verb*. That is, even if we recognize an array of NP-V-NP-*to*/*for*-NP as a single unit, it is not necessarily the example of the abstract structure of (3). Look at another example which exemplifies the pattern NP-V-NP-*to*/*for*-NP:

(5) That seemed the best choice to me. (c.f., That gave the best choice to me.)

This is indeed an example of NP-V-NP-*to*/*for*-NP, but is not considered as PDC. Perhaps the reason for the denial is related to the fact that the noun phrase "the best choice" is similar to adjective phrases in distribution. In fact, it cannot be replaced with a pronoun. Thus a sentence just below is not an acceptable one.

(6) *That seemed it to me. (c.f., That gave it to me.)[1]

Therefore, it can be said that "the best choice" in (5) is a noun phrase as a token, but not as a type. Then, what factor determines the type? Perhaps, it is a type of the verb within the pattern that determines the type of noun phrase following the verb (in fact, we can identify as PDC the sentences presented within parentheses in the sentences (5) and (6)), and the type in question is, in a word, *DA-Verb*.

Consequently, we cannot obtain any cue to identify PDCs only from the structural definition of (3). We need, therefore, more concrete cues which present us what lexical items can exemplify the pattern "DA-Subj DA-Verb DA-Obj *to* DA-Loc / DA-Subj DA-Verb DA-Obj *for* DA-Benef (= (3))." We will call this, henceforth, "linking problem" of an abstract structural description to concrete arrays of phrases (see fig. 1).

Semantic representation	HUMAN	ACT	THING(S)		LOCATION
Structural description	DA-Subj	DA-Verb	DA-Obj	*to*	DA-Loc
Categorical representation	NP	V	NP	*to*	NP
Concrete items (phrases)	*John*	*gave*	*the book*	*to*	*me*

fig. 1: linking image of PDC

[1] Asterisk on the top of the sentence (*) denotes the following sentence is not grammatical.

3. Theoretical background

Then, how can we solve the linking problem? In other words, how can we specify what *DA-Verb* is? This is an empirical question. We need a language theory or model which assures that a certain empirical method can specify the human knowledge about syntactic patterns, by which, for example, we can identify an array of phrases NP-V-NP-*to*/*for*-NP with a particular syntactic pattern "DA-Subj DA-Verb DA-Obj *to* DA-Loc / DA-Subj DA-Verb DA-Obj *for* DA-Benef (= (3))."

3.1. About Usage-based Model of Language (UML)

The leading candidate for such a model of language is Usage-based Model of Language (UML, hereinafter: e.g., Langacker 1987, 1991; Kemmer & Barlow 2000; Bybee 2001; Croft 2002; Tomasello 2003).

UML was advocated by the cognitive linguist Ronald Langacker in the program named as Cognitive Grammar (Langacker 1987, 1991). He established a theoretical background of UML, though he did not analyze certain linguistic phenomena (e.g., morphophonological structure of English, inflection systems of English verb) empirically.

UML is characterized as a radically inductive approach, as opposed to a highly deductive one such as Generative Grammar (e.g., Chomsky 1965). In UML, language is not regarded as an abstract rule-based system (e.g., Generative Grammar) but as a mass of concrete usages. Usage mass is considered to be a vast network which is self-organized by means of statistical structures underlying input information. Therefore, frequency is the most important factor. Language structure is assumed to emerge from language use (Bybee 2001). Some researchers argue that, in order to acquire a language, children only need abilities to understand other person's intention and to find recurrent patterns from the stream of sounds (e.g., Tomsello 2003) while others emphasize statistical processing ability which enable them to internalize the structure of input linguistic data (e.g., Haryu & Imai 2000).

3.2. Examples of UML

UML is embodied by several linguists. Joan Bybee applies UML to English morphophonological structures (e.g., Bybee 2001). She analyzes the English morphophonological system as an "associative network" (Bybee 2001: 23), in which morphemes are linked by semantic and phonological similarities. The network model can deal correctly with the irregularity of English morphological system (e.g., regular vs. irregular inflection of verb, nominalization of verbs) as a frequency-based self-organized structure.

UML is also applied to the field of language acquisition. Michael Tomasello voices his stance of "usage-based linguistics" in his famous book whose subtitle is *A usage-based theory of language acquisition* (Tomasello 2003). He constructs an overview of language structure by referring vast amount of data gained from several empirical researches including experiments and corpus analyses. He insists that the domain-general (socio-)cognitive abilities enable us to acquire a language. Those abilities are our "intention-reading ability" and "pattern finding ability" (Tomasello 2003: 3-4); these two are integrated into a powerful mechanism which makes indispensable such an innate faculty as universal grammar.

3.3. What does UML enable us to say?

Taking a usage-based view of language, we obtain a theoretical backing for statistical approach to syntactic structures. Researches within UML demonstrate that statistical information is crucial to language acquisition and construction. Therefore, it is highly likely to be valid to identify statistical descriptions with syntactic descriptions. In other words, it can be said that, rather radically, syntax is statistics (c.f., Hunston & Francis 2000).

4. Research

In this section, the result of a statistical analysis of English PDC is presented in order to demonstrate that statistical information can describe syntactic patterns.

4.1. Hypothesis

From a usage-based view, we can say that a certain syntactic pattern is perceived by us because of its statistical property. In other words, elements which consist of a syntactic pattern co-occur frequently enough for us to recognize that they form a pattern.

Then, as to PDC, we can frame a group of hypotheses as follows:

(7) a. It is because a phrasal pattern NP-V-NP-*to*/*for*-NP occurs frequently enough that we can find it.
b. It is because some verbs and the preposition *to* or *for* in the context NP-V-NP-*to*/*for*-NP co-occur frequently enough that we can perceive the whole pattern as an interdependent unit.

The former hypothesis is related to the fact that we can actually find the syntactic pattern in question and the latter the fact that we can identify a certain occurrence of the pattern with an example of PDC.

4.2. Method

In order to verify the hypotheses, we should design an appropriate research method. This research was conducted using the following procedure:

I. Examples of V-Pro-*to*/*for*-Pro[2] were collected from the English balanced corpus *British National Corpus* (BNC), which is annotated with part-of-speech (POS) tags.
II. Retrieval and sorting of examples (that is, concordance) were conducted using the on-line tool Sketch Engine (http://www.sketchengine.co.uk/) because Sketch Engine allows us to use regular expressions.
III. The retrieved data were statistically analyzed with the use of Microsoft Excel, in which the co-occurrence strength of verbs and *to* or *for* was calculated.

Note that in phase I, the retrieved pattern is not NP-V-NP-*to*/*for*-NP but V-Pro-*to*/*for*-Pro. There are two reasons for this: first, we would like to examine the examples forming infinitive phrases such as "to give the book to him," which is never found if we retrieve the pattern NP-V-NP-*to*/*for*-NP; second, noun phrase (NP) is highly difficult to retrieve from the corpus annotated only with POS tags, and pronouns and Proper Name is assumed to be the same as NP in distribution.

In the phase III, co-occurrence strength was measured by *t-score*, which is the statistical criterion showing "the degree of certainty that two words co-occur with greater than a chance probability" (Hunston & Francis 2000: 231). It is calculated as follows (f_c denotes the frequency of co-occurrence, f_a and f_b denotes the frequencies of each word, and n denotes the total size of corpus):

(7) $t = (f_c - (f_a \cdot f_b / n)) / \sqrt{f_c}$

This equation denotes that when we calculate t-score, 1) we multiply f_a by f_b; 2) the product of 1) is divided by n; 3) we subtract the quotient of 2) from f_c; 4) the result obtained thorough 1)-3) is divided by the square root of f_c. Suppose we calculate the t-score of the phrase *academic writing* in BNC. The frequency of *academic writing*, *academic*, and *writing*, and the total size of BNC are 21, 4614, 5236, and 111173004, respectively, so the t-score of

[2] *Pro* denotes Pronouns (both definite and indefinite) and Proper Nouns.

academic and *writing* = (21 − (4614 · 5236 / 111173004)) / √21 = (21-0.217) / 4.583 ≈ 4.535. In general, when t-score is equal to or more than two, co-occurrence of the words is considered frequent (Barnbrook 1996: 98; Hunston 2002: 72). Therefore, it can be said that *academic* and *writing* co-occur frequently in that order.

3.2. Results

As for V-Pro-*to*-Pro, 6080 examples were found. There were 391 verbs in the context. The most frequent verb was *give*, whose frequency was 834 (13.72%). The number of verbs whose t-score (with *to*) is equal to or more than two is 108 (27.671%).

As for V-Pro-*for*-Pro, 1777 examples were retrieved, which includes 414 verbs. The most frequent verb was *get*, whose frequency was 142 (7.99%). The number of verbs whose t-score (with *for*) is equal to or more than two is 71 (17.150%).

The average of t-score between verbs and *to* is 2.087 (more than two), but that between verbs and *for* is 1.470 (less than two). This contrast is discussed in the subsection just below.

Before discussing the data, let us look at the fragments of the total data obtained (see table. 1 and table. 2 just below).

	Verb	Freq.	Freq. rate	Freq. of V	t-score
1	give	834	13.717%	123424	28.65
2	take	441	7.253%	173808	20.55
3	say	380	6.250%	317301	18.60
4	send	311	5.115%	24186	17.56
5	leave	264	4.342%	63829	16.03
6	mean	219	3.602%	68783	14.54
7	hand	215	3.536%	55331	14.46
8	bring	191	3.141%	42478	13.65
9	introduce	158	2.599%	14247	12.51
10	show	131	2.155%	62324	11.15
11	keep	130	2.138%	48765	11.17
12	explain	128	2.105%	18610	11.22
13	sell	125	2.056%	21107	11.08
14	return	99	1.628%	23129	9.82
15	mention	96	1.579%	12333	9.73
16	owe	89	1.464%	3604	9.41
17	put	85	1.398%	67749	8.82
18	offer	67	1.102%	29985	7.99
19	get	63	1.036%	213313	6.47
20	read	55	0.905%	27975	7.21
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
391	yoke	1	0.016%	154	0.99
	total	6080	100.000%	average	2.09

table. 1: fragment of the data in V-Pro-*to*-Pro

	Verb	Freq.	Freq. rate	Freq. of V	t-score
1	get	142	7.991%	213313	11.63
2	leave	100	5.627%	63829	9.90
3	thank	62	3.489%	12639	7.85
4	ask	50	2.814%	57766	6.94
5	buy	50	2.814%	25503	7.01
6	find	48	2.701%	95923	6.71
7	blame	46	2.589%	5035	6.77
8	pay	40	2.251%	38830	6.23
9	make	38	2.138%	210744	5.62
10	want	37	2.082%	90471	5.85
11	keep	35	1.970%	48765	5.78
12	take	33	1.857%	173808	5.26
13	see	32	1.801%	185245	5.13
14	feel	31	1.745%	59764	5.39
15	write	28	1.576%	39335	5.17
16	hate	24	1.351%	5058	4.88
17	hold	21	1.182%	49668	4.41
18	give	18	1.013%	123424	3.78
19	use	17	0.957%	109238	3.70
20	mistake	15	0.844%	6220	3.85
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
414	wreck	1	0.056%	1099	0.98
	total	1777	100.000%	average	1.47

table. 2: fragment of the data in V-Pro-*for*-Pro

3.3. Discussion

We will consider what is implied by the fact that the average t-score of V-Pro-*to*-Pro is over two while that of V-Pro-*for*-Pro is not. Probably this implies that the phrasal chain

V-Pro-*to*-Pro can cause us to perceive it as a single interdependent pattern but V-Pro-*for*-Pro itself cannot. Therefore, when we encounter a certain example of V-Pro-*to*-Pro (e.g., *give it to you*), we can recognize it as a syntactic pattern, that is, PDC; on the other hand, even if we come across a certain sequence of V-Pro-*for*-Pro, we cannot regard it as a single syntactic pattern.

This contrast is probably due to the fact that the preposition *for* has a stronger meaning than *to*. It can be said that *for NP* itself can imply "for the sake of *NP*" or the like, but *to NP* does not have any independent interpretation. This means that interpretation of *to NP* is highly dependent on the occurrence context such as V-NP-*to*-NP. In fact, Tomasello (2003) also reports that young children use *for*-dative constructions with more various verbs than *to*-dative constructions and suggest that this asymmetry is due to the usability of *for*-dative constructions in any situation in which an action is of benefit to someone.

Considering the fact that the average of t-score between verbs and *for* in the context V-Pro-*for*-Pro is less than two, one can conclude that NP-V-NP-*for*-NP is not a single syntactic unit (or, construction) but a composite of two or more elements. However, there is a factor which makes us think it is a syntactic pattern. The factor is a "convertibility of construction," that is, we can convert some NP_1-V-NP_2-*for*-NP_3 to NP_1-V-NP_3-NP_2:

(8) a. John bought the book for me.
    b. John bought me the book.

This convertibility is also found in *to*-dative construction:

(9) a. John gave the book to me.
    b. John gave me the book.

It is assumed that the high strength of co-occurrence and the convertibility are integrated to cause us to perceive the sequence as a single syntactic unit.

4. Conclusion

In conclusion, we can say that V-Pro-*to*-Pro is perceived as a syntactic unit because *V* and *to* in the context co-occurs frequently enough. In other words, high frequency links a categorical sequence with abstract structural representation.

However, it is not necessarily true that the obtained data reflect human cognition or knowledge about syntactic patterns. In order to assure the compatibility between statistical data and human cognition, we have to conduct some psychological experiment. Therefore, statistical analysis can only suggest, not demonstrate, that statistics descriptions represent syntactic structures. Incidentally, the positive correlations between statistical structures obtained by corpus analyses and behavioral data observed through psychological experiments are reported by some corpus linguists (e.g., Gries, Beate & Schönefeld 2005). Taking the results of such studies into consideration, the claim of this paper is though to be on the right track.

Bibliography

Barnbrook, Geoff. 1996. *Language and computers: A practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Campbell, Aimee L., and Michael Tomasello. 2001. The acquisition of English dative constructions. *Applied psycholinguistics* 22. 253-267

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge: MIT Press

Croft, William. 2002. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Goldberg, Adele. 1995. *Constructions: A construction grammar approach to argument structure.* Chicago; London: University of Chicago Press.

Gries, Stefan Th., Beate Mampe, and Doris Schönefeld. 2005. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive linguistics* 16 (4). 635-676

Haryu, Etsuko, and Imai, Mutsumi. 2000. Goi-gakushuu-mekanizumu ni okeru seiyaku no yakuwari to sono seitoku-sei [The role of constraints on the mechanism of lexical learning and its innateness]. *Kokoro no seitoku-sei* [*Innateness of mind*], ed. by Mutsumi Imai, 131-171. Tokyo: Kyoritsu Shuppan

Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, Susan, and Gill Francis. 2000. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Kemmer, Suzanne, and Michael Barlow. 2000. An introduction: Usage-based conception of language. *Usage-based models of language*, ed. by Suzanne Kemmer and Michael. Barlow, vii-xxii. Stanford: CSLI Publications.

Langacker, Ronald. 1987. *Foundations of cognitive grammar Vol. 1: Theoretical prerequisites*. Stanford: Stanford University Press.

——. 1991. *Foundations of cognitive grammar Vol. 2: Descriptive application*. Stanford: Stanford University Press.

Newman, John. 1996. *Give: A cognitive linguistic study.* Berlin: Mouton deGruyter.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Harvard: Harvard University Press.