半具体データからの抽象的な文法特性の抽出:
多変量解析による英語高頻度動詞の抽象化の試み

Extracting abstract properties from the semi-concrete data: Clustering English high-frequency verbs by multivariate analysis.

吉川正人[†]
Masato YOSHIKAWA

[†]慶應義塾大学大学院
Keio University
machayoshikawa@dream.com

Abstract

In Usage-based Model of language, the reason for the existence of abstract grammatical properties describable as a set of rules is not attributed to the character of language itself as the system of rules, but to "emergence" of such properties from the large body of declarative linguistic knowledge. This assumption is, however, yet to be validated empirically on a large scale.

Therefore, this paper is aimed at presenting a large-scale corpus-based statistical research for the English syntactic structure in which high-frequency verbs are clustered based on the semi-concrete data of co-occurrences with items in any kind of grammatical relations (such as subject, object, modifier, and so on).

The result of this research suggests that the semi-concrete co-occurrence data can successfully guide a language learner to abstract grammatical knowledge of English, which supports the assumption of Usage-based Model of language.

Keywords — Usage-based Model, Word Sketch, multivariate analysis, WordNet, English syntax.

1. はじめに

言語の用法基盤モデル(Usage-based Model [4])では、自然言語に抽象規則として記述可能な文法特性が備わっているのは、宣言的な言語知識からそのような特性が創発するからだと主張される。しかしこの想定の実証(e.g., [1][2])は依然断片的なものであり、大規模な実証研究が求められる。そこで本稿では、大規模コーパスを用いて、英語の高頻度動詞をクラスタリングし、抽象的な文法特性が(半)具体的な共起情報からある程度指定可能であることを示す。具体的には、以下のようなコーパス調査を行う:

(1) a. 英語の均衡コーパスである British National Corpus (BNC)から、
b. Sketch Engine [5]上のツール Word Sketch を用いて、
c. 高頻度動詞(BNC で生起回数が 1 万回以上の動詞)に対し、
d. 主語や目的語､修飾語となっている語とその頻度を抽出し、
e. c の動詞に対し、d の情報に基づく多変量解析を行う

この調査の結果、このような共起データから、ある程度の統語特性の近似が行えることが明らかになった。

2. 調査

2.1. Word Sketch とは

Word Sketch(以下 WS)とは、ある語(動詞・名詞・形容詞に限られる)に対して、コーパスに付与された品詞タグ情報に基づいて規定された主語、目的語、修飾語等の文法関係にある語を、その生起頻度および統計スコアと共に表示するシステムである｡文法関係は品詞の配列に基づいて規定される。

WS の情報は、Python の API (http://trac.sketchengine.co.uk/wiki/SkE/Methods/ 参照)を用いて抽出した。調査対象となった動詞(=BNC で生起頻度が 10000 以上の動詞)は 216 語である。table. 1 に一部を提示する。

Table. 1: 調査対象となった動詞の一部

ID	verb	ID	verb	ID	verb	ID	verb
1	accept	38	continue	177	set	210	watch
2	achieve	39	control	178	share	211	wear
3	Act	40	cover	179	show	212	win
4	Add	41	create	180	sit	213	wish
5	admit	42	cut	181	smile	214	wonder
6	affect	43	deal	182	speak	215	work
7	agree	44	decide	183	spend	216	write

2.2. 共起ベクトル

216 の各動詞に対し、WS から得られた文法関係の情報に基づく共起データを、1 アイテム 1 次元のベクトルとして与えた。例えば、動詞 *write* とは *article* が目的語として 263 回共起しているが、この場合、“object-article”という次元に“263” という値が与えられる。この際、生起頻度が 10 を下回るものは排除し、結果 15,884 次元のベクトル表示を得た。

このベクトル上の類似度を元に、動詞の多変量解析を行った。概念的には、同じ文法関係に同じ語が生起している数が多ければ多いほど、その動詞は似ていると見做されるということになる。手法としては、主成分分析を用いた。

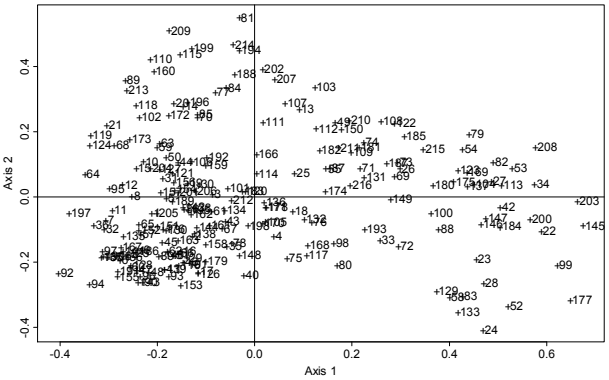
2.3. 主成分分析

主成分分析は、Microsoft Excel の“correl”関数を用いて作成した動詞の相関行列を入力とし、Rgui (version 2.7.1)を用いて行われた。成分は、分析の容易さを考えて、2 次元に圧縮された。

3. 結果と考察

主成分分析の結果を散布図にして提示する(fig. 1; 散布図は分析同様 Rgui のプログラムを用いて行われた)。

Fig. 1: 主成分分析の結果 (数値は各動詞の ID)



この結果は以下のように評価された: 1) WordNet [3]の上位語データを利用して動詞を分類し、2) その分類と散布図上の I~IV 象限への分布とに対応が見られるかどうか検証した。各動詞に一律に深さ 2 の上位語(=トップノードの一つ下)を与えた。語義が複数あり上位語にも複数の候補が存在する場合は、一つ目の上位語を割り当てた。

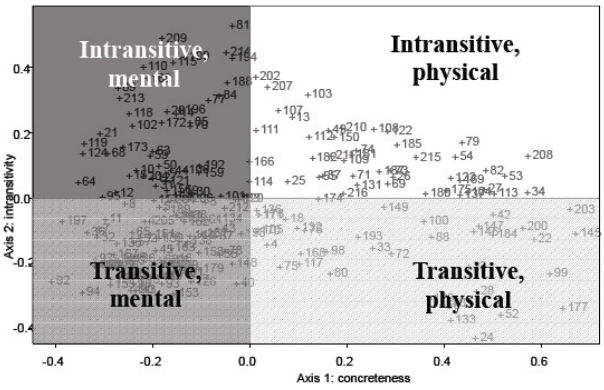
結果、各象限にそれぞれ特異的な上位語が現れることが明らかになった。それぞれの象限に現れた特徴的な動詞上位語を table. 2 に提示する:

Table. 2: 各象限に特異な動詞上位語

象限	特徴的な上位語	動詞の事例
I	TRAVEL	*arrive, come, drive, drop, fall, get, go, leave, move, run, walk,* etc
II	THINK	*assume, believe, expect, feel, hope, know, remember, suppose, think,* etc
III	ACT, CHANGE, MAKE	*accept, admit, agree, cause, compare, enable, encourage, note, plan,* etc
IV	MOVE	*break, bring, build, cut, fill, give, grow, keep, open, put, throw,* etc

ここから、主成分 1(=X 軸)を具体性、主成分 2(=Y 軸)を自動性の次元として解釈可能であると言える。具体性とは、統語的には前置詞句との共起および具体名詞との共起という形で実現されているようである。図示すると fig. 2 のようになる。

Fig. 2: 各象限の特徴づけ



4. 結語

以上より、(半)具体的な共起データから、ある種の抽象的な文法特性が抽出可能であることが示唆される。尤も、本調査の結果は抽象的な文法特性へのかなり荒い近似を示したまでであって、より複雑な振る舞いまで具体データから規定するにはより精緻な調査が必要になるのは言うまでもない。

参考文献

[1] Elman, J, (1990) “Finding structure in time”, *Cognitive science, 14,* pp. 179-211.

[2] ——, (1991) “Distributed representations, simple recurrent networks, and grammatical structure”, *Machine learning, 7*, pp. 195-225.

[3] Fellbaum, C, (1998) *WordNet : An electronic lexical database.* Cambridge: MIT Press.

[4] Kemmer, S., & Barlow, M, (2000) “Introduction: A usage-based conception of language”, In Barlow, M., &. Kemmer, S. (eds.) *Usage-based models of language* (pp. vii-xxii). Stanford: CSLI Publications.

[5] Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D, (2004) “The Sketch Engine”, *Proceedings of EURALEX*, pp. 105-116.