様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　24 年　6 月　1 日現在

機関番号：34310
研究種目：若手研究（Ｂ）
研究期間：2009～2011
課題番号：21720179
研究課題名（和文）　記号主義言語論に基づく英語動詞構文ネットワークのデータベース化

研究課題名（英文）　An Attempt to Build a Database of Basic English Verb Constructions based on Symbolic View of Language

研究代表者
長谷部　陽一郎（HASEBE YOICHIRO）
同志社大学・グローバル・コミュニケーション学部・准教授
研究者番号：90353135

研究成果の概要（和文）：構文ネットワークの構造が記号主義的言語観のもとに、どのように表現可能であるかを検討した上で、英語基本動詞の構文ネットワークを計算機上のデータベースとして実装する試みを行った。コーパスやウェブから得られたデータを用いたシステムを実際に構築することにより、理論の妥当性を示すと共に、言語教育や自然言語処理にも有用な言語資源の構築方法を確立した。

研究成果の概要（英文）：This research project aimed at elucidating the way linguistic constructions form a network, and how such a network should be represented. It also tried to build an actual database system of linguistic constructions involving fundamental verbs of English. The resulting system is capable of visualizing complex relationships among verbs and their constructions. The method that made it possible will offer not only a proof of concept, but also useful suggestions from theoretical linguistics for projects in related areas such as language education and natural language processing.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	700,000	210,000	910,000
２０１０年度	500,000	150,000	650,000
２０１１年度	500,000	150,000	650,000
総　計	1,700,000	370,000	2,210,000

研究分野：人文学
科研費の分科・細目：言語学
キーワード：語彙・意味

１．研究開始当初の背景

（１）「記号主義的言語論」とは、語や句や文を含む、あらゆる言語構造が「意味」と「形式」との記号的関係に基づくと考える理論言語学上の潮流である。記号主義的言語論の歴史は近代言語学の祖、ソシュールにまで遡るが、近年においても、Charles Fillmore、Adele Goldberg らによって進められている構文文法理論や、Ronald Langacker の認知文法理論などが記号主義を明確に打ち出している。

これら現代の記号主義的言語論の特徴の一つは、「構文」という概念を重視することである。様々なレベルの構文における記号関係を明らかにすることは、自然言語のメカニズム解明に不可欠とみなされている。また、構文はネットワーク的構造の中に位置付けられており、互いに継承、汎化、拡張といった関係で結ばれていると考えられている。

しかし、そのように重要な「構文」の概念について、計算的ないしは数学的に厳密な定義は必ずしも示されていない。また、「ネットワーク」の形状についても、漠然とした「イ

メージ図」ないしは「スキーマ図」といったものの他に、構造を視覚的に示すための理論や方法は確立されていない。本研究は、構文文法や認知文法といった記号主義的言語論の基本的な正しさを認識しつつ、上記の課題に対して向き合う一つの試みとして構想された。

（2）従来、言語教育や自然言語処理等の周辺分野では、理論言語学の分野での知見は言語に対する単なる「見方」や「捉え方」を示すに過ぎず、実際に応用可能な「システム」を提供しない、という批判がなされてきた。かつて理論言語学研究は、現象を正しく記述し、そこからある種の一般化を行うことが基本的な仕事であった。しかし現代では教育や情報工学をはじめ、様々な領域で言語データを効果的に扱い得る知見が必要とされており、理論言語学には何ができるかを具体的に検討することが求められている。本研究では、応用研究を必ずしも想定していない言語理論に基づきながらも、実際的な機能を持ったデータベース・システムの構築を試みた。

２．研究の目的

（1）本研究の大まかな目的は 1 で示した通りであるが、具体的には次のことを目指した。まず、「意味」と「形式」の結びつきがネットワークを形作るという考え方を、どのように計算機上に実装するか、その手法の確立である。本研究の開始以前より、形式概念分析（Formal Concept Analysis = FCA）と呼ばれる手法が有用であろうという予測はあった。しかし、語や句や構文といった言語単位をはじめ、理論言語学上の様々な概念をいかにして FCA の手法に落とし込んでいくかは明確でなく、この問題に取り組むことは大きな課題であった。

（2）FCA に基づいて、意味と形式の結びつきとしての構文と、構文間のネットワーク構造とを計算機的に表現することが第 2 の目的である。もちろん、それは自然言語を操る話者の知識や言語運用能力を完全に再現するということではない。あくまで理論的な記述を計算機上に移し替えるに過ぎない。しかし、そこには数学的、論理的な厳密さが求められるため、このプロセスを経ることで、もともとの理論の妥当性や整合性の度合いを測ることができる。

（3）動詞構文のネットワークをデータベース・システムとして計算機的に実装する際のモデリングが実際の言語知識に近いものであるなら、そのようなシステムは言語教育や自然言語処理においても何らかの形で有用なものになると予想される。言語の理論的研究の成果を応用領域において活かすための可能な方法を探ることも本研究の目的の 1 つである。

３．研究の方法

（1）<u>方法論的検討</u>　構文がネットワーク関係を構築するならば、要素となる構文の意味的側面と形式的側面とが様々な度合いで結びつきながら、継承、汎化、拡張といった関係を構築すると考えられる。このように複雑なネットワーク構造を記述ないしは分析するための数学的手法として FCA がある。本研究の前半では、FCA を用いて構文ネットワークを表現するための実験を行った。その際、まずは英語のすべての構文ではなく、様々な理論の中で盛んに研究されてきた「動詞構文」のバリエーション（NP1 + V + NP2, NP1 + V + NP2 + NP3, etc.）に的を絞り、いかに要素を FCA に落とし込んでいくかを検討した。これにより、FCA を用いた動詞構文のネットワークを記述する方法が有効であることが確認された。

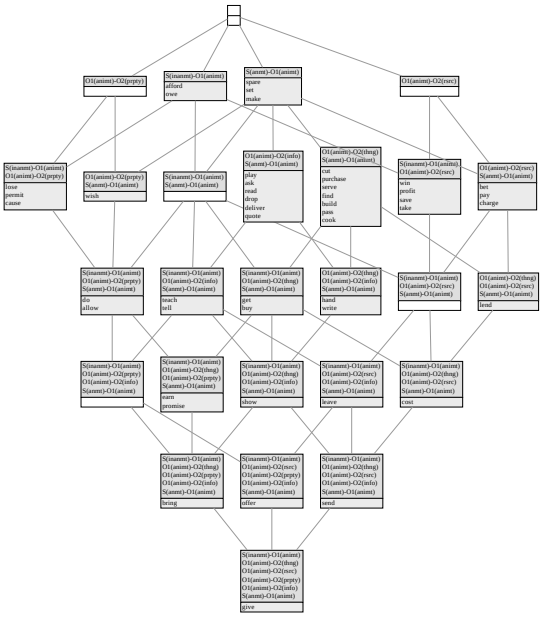

図 1

図 1 は、英語他動詞構文の意味的側面に関する情報を項名詞句が取る意味役割の組み合わせとして設定し、FCA の処理を行った結果である。紙面の都合上、詳細を解説することはできないが、意味と形式（=ここでは NP1 + V + NP2, NP1 + V + NP2 + NP3, etc.といった統語的表示）の複合的なパターンが記号的性質を備えた「構文」のあり方を（部分的ではあっても）適切に表現している。

しかしながら、研究を重ねる中で、上記の方法がスケーラビリティの面で必ずしも最適ではないことが次第に明らかとなった。す

なわち、他動詞構文のように対象を限定した場合には有効であるが、多くの構文を包括的に扱おうとした際には必ずしも上手くいかないのである。FCA では対象構造が持つ 2 つの側面について、各々を有限個の特徴ラベルの組み合わせとして表現する。しかしあらゆる項名詞句の意味役割を限られた数のラベルで表現することは非常に困難な作業となる。ラベルが明確に規定できない場合、FCA の有効性は期待できない。

結果として、項名詞句の意味役割の組み合わせにより動詞の意味カテゴリーを間接的に規定する方法は破棄することとなった。実際のところ、名詞はすべての品詞の中で最も「オープン」なクラスであり、それ自体が高い多態性を有している。そのような名詞の意味役割記述を「厳密に」行うことは現実的ではない。そこで新たに目を付けたのが、動詞と同じ節に共起する副詞の持つ意味カテゴリーである。副詞もまたオープンなクラスであるが、異なり語数においても、また意味的なカテゴリーの数においても、名詞に比べると圧倒的に規模が小さい。このように共起副詞のカテゴリー情報を用いることで、広汎な動詞構文のネットワーク構造を計算機上で構築することが可能になった。

（2）基本データの取得　本研究の目的の 1 つは、実際のデータを格納した構文データベースを計算機上のシステムとして実装することであり、そのためには必要な言語データを採取し、適宜、分類・加工・調整することが必要であった。

本研究のシステムに入力する基本動詞を選定するにあたっては、British National Corpus の頻度表（1 位から 5000 位まで）を取得し、その中から BNC の品詞解析の不備によるとみられるノイズ・データを削除し、さらに複数の既存語彙リストに収録されている語群を参考に、3883 語から成るリストを作成した。

次に構文ネットワークの「意味的側面」を担う共起副詞は、BNC や Wikipedia ダンプファイルを基にしたデータを中心に採取し、加えてこれらコーパスの内容的片寄りから生じるとみられる情報欠損を補うため、複数の辞書、辞典を用いて、当該の動詞と共起し得る副詞のデータを蓄積した。その数は異なり語数で 3854 語であった。

動詞構文の「形式的側面」を担う統語パターンについては、Princeton 大学で開発され、一般公開されている語彙情報データベース WordNet のデータを用いた。WordNet では動詞が取り得る統語パターン（=フレーム）を 35 種に分類しており、それぞれの動詞で利用可能なパターンの情報が得られるようになっている。例えば動詞 ***help*** であれば、[Somebody ----s somebody]や[It ----s that CLAUSE]といった、19 のフレームが関連付けられている。これらを利用することで、基本動詞の形式的側面について、かなり網羅的な記述が可能になる。しかし合計で 35 種のフレームというのは、FCA 処理に落とし込むにはやや多すぎる数である。フレームの数が大きくなっている要因として、WordNet では項名詞句の有生性を弁別しているということがある。そのため例えば、[Somebody ----s]と[Something ----s]とは別のフレームとして扱われる。しかし、有生／生物の区別は、認知言語学における概念メタファーの重要性などを考え合わせると、揺らぎのない厳密な区別と言えない。また、そもそも意味と形式の結びつきを中心に据えた立場からすると、形式構造（NP1 + V + NP2）に意味役割（somebody / something）を入り込ませた見立ては妥当と言えない。そこで本研究では項の有生／無生の別を排除して表 1 のように 27 個のフレームにまとめたデータを利用することにした。

表 1　27 種のフレーム・データ

ID	フレーム形態
1	A ----s
2	It is ----ing
3	A is ----ing PP
4	A ----s B Adjective/Noun
5	A ----s Adjective/Noun
6	A ----s Adjective
7	A ----s B
8	A ----s to B
9	A ----s on B
10	A ----s B C
11	A ----s B to C
12	A ----s B from C
13	A ----s B with C
14	A ----s B of C
15	A ----s B on C
16	A ----s B PP
17	A ----s PP
18	A's (body part) ----s
19	A ----s B to INFINITIVE
20	A ----s B INFINITIVE
21	A ----s that CLAUSE
22	A ----s to INFINITIVE
23	A ----s whether INFINITIVE
24	A ----s B into V-ing C
25	A ----s INFINITIVE
26	A ----s VERB-ing
27	It ----s that CLAUSE

（3）動詞の特徴的属性の抽出　上記の作業で得られた副詞データを用いて動詞構文の意味的側面（傾向）を間接的に規定する試み

を行った。ここで 3854 の異なり語をすべて個別の要素とみなして FCA 処理にかけることはできない。これらの副詞はそのままでは単なる「共起語」であり、そこから何らかの「意味」を見出すためには人間的な作業が必要となる。そこで行ったのは、副詞の意味タグのセットを設定し、すべての副詞にタグを 1 個以上付与することである。理論上、副詞の意味構造は述語動詞句のそれに融合されるため、共起副詞の意味カテゴリーの総和を取ることで、動詞のスキーマ的意味構造を（間接的・部分的ではあるが）規定することができる。

副詞に付与する意味カテゴリーに関しては、Geoffrey Pullum と Rodney Huddleston が *The Cambridge Grammar of the English Language* において示したものが良く知られている。すなわちそれは、副詞を manner, act-related, spatial location, temporal location, duration, aspectuality, frequency, serial order, degree, reason, concession, condition, domain, modality, evaluation, speech act-related というカテゴリーに分類したものである。しかし、副詞の中には文修飾副詞と呼ばれるものもあり、本研究の目的にとってそれらは不要である。また、副詞の形容詞修飾の機能についても考慮する必要がない。そこで、上記の意味カテゴリーを選別・修正し、本研究では表 2 に示す 10 の意味カテゴリー・タグを用いることにした。

表 2　動詞修飾副詞の意味カテゴリー・タグ

ID	意味カテゴリー・タグ
1	TEMP/ASPECT
2	LOCATION
3	MANNER
4	SUBJECTIVE
5	DEGREE
6	DOMAIN
7	ORDER
8	DURATION
9	FREQUENCY
10	MEANS

これらの意味カテゴリー・タグを 3854 語の副詞に対して付与する作業は手作業で行った。語によっては単一のタグが付与されるが、多くの場合、複数のタグが付与される。例えば、*loudly* であれば、MANNER タグのみであるが、*happily* であれば、MANNER タグに加えて SUBJECTIVE タグが付与される。あるいは *impressively* であれば、MANNER、SUBJECTIVE に加え、DEGREE タグが付与される。（なお、SUBJECTIVE は意味として表示する内容に認知主体の主観的な評価・判断が色濃く反映されるタイプの副詞に対応するカテゴリー・タグである。）

こうして付与されたタグのパターンは複数の副詞の間で一致しており、それ自体は各副詞を特徴付けるものではない。この作業の目的はあくまで、副詞と共起する動詞の意味タイプを間接的に規定することである。ある動詞と共起する副詞のリストを得て、それらに付与された意味カテゴリー・タグを集計すると、動詞の大まかな意味タイプが浮かび上がってくる。しかし、本研究で用いた共起副詞データは、コーパスから取得し、既存の辞書・辞典等から得た情報で調整を行ったものである。その中で、高頻度な動詞は長い共起副詞リストを伴っているが、逆に低頻度語にはわずかな数の副詞のみが与えられている。このような出現頻度による情報量の差は、意味タイプの推定を行う際に問題となる。

これを低減するための措置として、共起副詞を通じて結びついた意味カテゴリー・タグの TF-IDF 値を計算し、その結果から各動詞にとって「特徴的」な意味カテゴリー・タグを抽出した。TF-IDF とは文書中に出現する単語の重みを得るための手法として情報検索や文書要約の分野で広く利用されている。本研究ではこれを応用し、動詞を「文書」、共起副詞の意味カテゴリー・タグを「単語」とみなして計算を行うことで、動詞のいわばスキーマ的意味の推定を試みた。実際に使用した計算式は次の通りである。

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = 当該の動詞 j における意味タグ i の数
df_i = i を含む動詞の総数
N = 全動詞数

TF-IDF の値が大きいほど、「文書」と「単語」の結びつきは強いとみなされるため、高い TF-IDF 値が算出された意味タグはその動詞の「特徴的属性」として扱うことができる。TF-IDF では、データ中に広く出現する要素の値は低く、特定の「文書」（=ここでは動詞）の中で現れる要素の値は高くなる。そこで、ほとんどの動詞と共起するような副詞に付与されたタグについては、当該動詞に対して圧倒的に強く関連付いているのでない限り高い TF-IDF 値につながらず、特徴的属性となりにくい。

このようにして共起副詞から得られた動詞の意味属性に順列を付け、全体の TF-IDF 平均値を閾値とすることで、各動詞の特徴属性を決定した。例えば動詞 *kiss* であれば、属性と TF-IDF 値のペアは表 3 のように示される。本研究における全動詞属性 TF-IDF の平均は 5.19 であり、これを閾値として決定された動詞 kiss の特徴属性は MANNER, SUBJECTIVE, DEGREE, TEMP/ASPECT の 4 種（網掛部）となる。なお、ここで TF-IDF

の値が低い属性は動詞 *kiss* と必ずしも「共起しにくい」または「馴染みにくい」わけではない。なぜならこれらの属性を持った副詞との共起も実際に起こっているからである。しかし TF-IDF の性質に鑑みて、これらの属性は、当該動詞がその特性上共起を強く求めるようなものではないと言える。

表 3　動詞 *kiss* の意味属性

意味属性	TF-IDF
MANNER	24.0
SUBJECTIVE	16.0
DEGREE	11.0
TEMP/ASPECT	6.0
DURATION	3.4
FREQUENCY	3.0
LOCATION	3.0
ORDER	3.0
DOMAIN	2.1

（4）FCA によるネットワーク構造の可視化

以上の取り組みにより得られたのは、3883 の動詞がそれぞれ取り得る統語パターンの組み合わせ（表 1 参照）と、高い TF-IDF を示す特徴属性(表 2 参照)の組み合わせである。特定の動詞を指定し、統語パターンと特徴属性のリストを FCA 処理にかけることで、その動詞が展開する構文（動詞構文）とネットワーク構造、そしてネットワークの要素となる各構文を展開する他の動詞のリストを出力することができる。記号主義的言語観のもとに形式と意味の側面から構文ネットワークを構築するという本研究の目的により、それぞれの側面に着目する形で結果を出力すると、図 2 および図 3 のようになる。これらは例として動詞 *dream* の動詞構文ネットワークを可視化したものである。

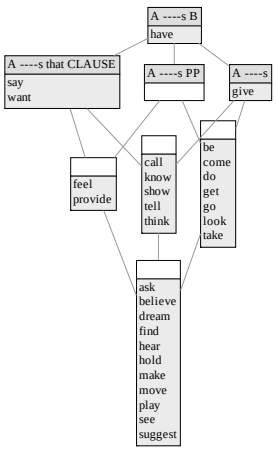

図 2　動詞 *dream* の構文ネットワークを形式の側面から見た図式

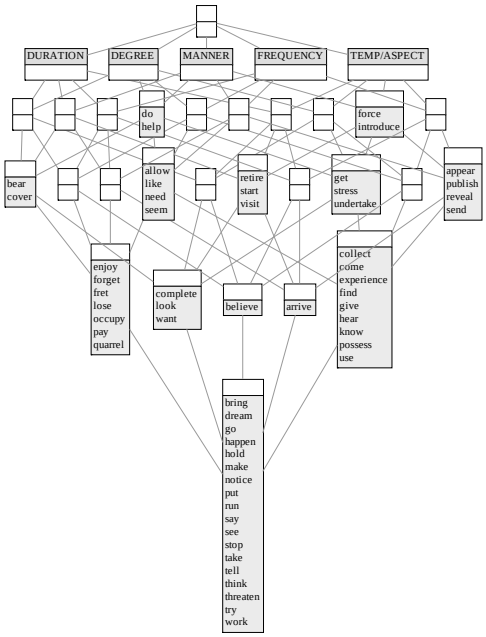

図 3　動詞 *dream* の構文ネットワークを意味の側面から見た図式

図の詳細な解説は紙面の都合上割愛するが、これらの図は基本的に、動詞 *dream* が取り得る形式的、意味的な現れの様々なパターンを広義での「構文」として捉え、そのネットワーク構造を示したものである。ネットワークの各構成要素が他の動詞からも参照されている場合は、当該動詞に加えて、それらの動詞のリストを表示している。構文文法理論において、動詞は「動詞構文」を構成し、動詞構文は同種の動詞構文と共に上位の「動詞クラス構文」を構成する。このような観点から考えると、図 2 および図 3 の中で動詞リストを含んでいる要素は、*dream* の動詞構文サブクラスであると同時に、動詞クラス構文の候補でもあると言える。

４．研究成果

本研究で得られた知見や技術の重要なものは３節で示した通りである。また、研究の過程における取り組みのいくつかは５節に示す論文等において発表済みである。その他、本研究では以下のようなシステムおよびツールを開発し、成果として公開している。

（1）ウェブインターフェイス・システム

本研究において構築されたデータベース・システムは、ウェブ・インターフェイスを通じて利用可能である。３節で示した構文ネットワークの可視化機能に加えて、各動詞の特徴属性と、それらの属性を付与する元となった共起副詞のリストを表示することができる。例えば、動詞 *kiss* を指定すると、意味属性の

うち、TF-IDF 値が高い 4 つの特徴属性と、それらに対応する共起副詞を表 4 のように表示する。

表 4　動詞 kiss の出力情報

特徴属性	共起副詞リスト
MANNER/ SUBJECTIVE （重複）	*gallantly, perfunctorily, fondly, dutifully, fervently, affectionately, tenderly*
DEGREE	*fervently, wildly, hard, lightly, really, well*
TEMP/ASPECT	*actually, always, just, only, now, then*

（2）データ処理ツールの公開　上記システムを構築するためには、いくつかのソフトウェア・ツールを改良または新規に作成する必要があった。中でも重要なのは、FCA に基づいたデータ処理を行い、ネットワーク図式を出力する RubyFCA プログラムと、Wikipedia ダンプデータなど言語的解析の行われてないコーパスデータ中の語に品詞タグ付けを行う EngTagger プログラムである。これらの言語処理ツール群は、英語教育や自然言語処理の分野での利用を想定し、一般公開している。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 6 件）

① 長谷部陽一郎・淺尾仁彦. 2011. 「PLM と FCA による構文ネットワークの記述について」『日本認知言語学会論文集 No.11』626-629. （査読有）
② 長谷部陽一郎. 2011. 「WordNet データベースと大規模コーパスからの頻度情報を用いた語彙概念習得支援ツール作成の試み」『2010 年度 ICT 授業実践報告書』55-64. （査読無）
③ 長谷部陽一郎. 2011.「第二言語語彙指導のための概念地図利用の可能性」『比較文化研究 No.96』 219-232. （査読有）
④長谷部陽一郎. 2010. 「構文のネットワークモデルについて—二重目的語構文を中心に」『認知言語学論考 No.9』81-137. （査読有）
⑤ 長谷部陽一郎. 2009. 「計算的手法を用いた構文習得の可能性」『言語文化』395-420. （査読有）
⑥ Yoichiro Hasebe and Kow Kuroda. 2009. “Extraction of English Ditransitive Constructions Using Formal Concept Analysis” *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation,* Vol.2. 678-685.（査読有）

〔学会発表〕（計 5 件）

① 長谷部陽一郎.「オブジェクト指向に基づく現実世界のモデリングと認知文法」第 1 回認知文法研究会（同志社大学） 2012 年 3 月 14 日.
② Yoichiro Hasebe. “A Corpus-based Approach to English Ditransitive Construction: The Causal Relation of Two Events” at The 11th International Conference of Cognitive Linguistics (Xi'an International Studies University, China) 2011 年 7 月 14 日.
③ 長谷部陽一郎・淺尾仁彦.「PLM と FCA による構文ネットワークの記述について」日本認知言語学会第 11 回全国大会（立教大学） 2010 年 9 月 11 日.
④ 長谷部陽一郎.「コーパスからの構文ネットワーク構築 : 英語他動詞構文の諸相」京都言語学コロキアム第 7 回年次大会（京都大学） 2010 年 8 月 28 日.
⑤ Yoichiro Hasebe and Kow Kuroda. “Extraction of English Ditransitive Constructions Using Formal Concept Analysis” at The 23rd Pacific Asia Conference on Language, Information and Computation (City University of Hong Kong, China) 2009 年 12 月 5 日.

〔その他〕
① データベース・システム公開サイト
http://yohasebe.com/verbmapr/
② 言語データ処理ツール公開サイト
http://github.com/yohasebe/

６．研究組織
(1)研究代表者

長谷部　陽一郎（HASEBE YOICHIRO）
同志社大学・グローバル・コミュニケーション学部・准教授
研究者番号：90353135