# 語彙概念構造を用いた日本語述語項構造コーパスの設計

松林 優一郎 *　　宮尾 祐介 *　　相澤 彰子 *§
* 国立情報学研究所コンテンツ科学研究系，§ 東京大学情報理工学系研究科
{y-matsu, yusuke, aizawa}@nii.ac.jp

## 1　序論

本稿では、現在我々が進めている、日本語の述語項構造タグ付けコーパスの開発状況について紹介する。

近年、文の統語構造が高精度で解析可能となったことを背景として、言い換えや含意関係認識などの処理のように、文の内外にある意味構造や、二つの言論の間の意味関係に注目した研究が盛んになってきた。これらの問題に対して鍵となる技術の一つに、述語項構造解析がある。これは、文中の述語と、その他の構成要素の間にある構文関係あるいは意味関係を認識する技術である。特に、言い換えや含意関係認識を実現するためには、異なる述語間の意味関係や、さらにはそれらの項構造の間の意味的関係を明らかにしなければならないことは明白であり、述語項構造解析は、この意味で重要なタスクである。

述語項構造解析に関する技術は、学習対象となるタグ付きコーパスの整備と共に発展してきた。英語圏では、FrameNet [9] や、PropBank [5]、VerbNet [6] 等を組み合わせたデータが、項構造の間の統語的／意味的関係を与える資源として機能した。一方で、日本語では、これらの関係を記述した十分な資源は未だ整備されていない。十分な量の解析済み述語項構造データを含んだ既存のコーパスとして、NAIST テキストコーパス [2] が挙げられるが、この資源は、ガ、ヲ、ニの三つの格助詞に対してのみ解析を与えており、述語の主要な項を網羅的におさえるのに不十分である。また、意味的な特性という意味では曖昧性を持っている格助詞をそのまま解析用のタグとして用いているため、各項の統語規則や意味機能を捉えて学習を行う必要のある、技術開発用のデータとしても、情報が不十分である。

我々の目的は、日本語の述語項構造解析の研究に必要とされるであろう豊富な意味情報を含んだ言語資源を構築することである。我々が開発するコーパスは、（1）各述語に対する項の完全なリストや、それらの項が持つ意味機能を表現する意味構造を含んだ、フレーム辞書と呼ばれる辞書と、（2）その辞書に基づいて述語項構造がタグ付けされた日本語文書の二つの部分からなる。述語項構造は、フレーム辞書の中で、語彙概念構造 (Lexical Conceptual Structure; LCS) [3] の理論に基づいて表現されている。LCS の理論では、述語の意味構造は複数の基本述語と呼ばれる述語の組み合わせによって分解される形で表現される。我々が LCS を用いる動機としては、この構造が、ある述語のそれぞれの統語的な項に対して明確な意味機能を規定する点があげられる。LCS 構造を構成する各々の基本述語が持つスロットは、大まかには従来の主題役割の意味機能に対応する。そうでありながら、LCS の利点は、各基本述語スロットの持つ意味が高度に機能化され、意味的な重複を持たない点にある。加えて、文中のある一つの統語的な項が、LCS 構造内の基本述語の複数の異なるスロットを同時に埋めることにより、各項がこれらの高度に汎化された意味機能のうち、複数の機能を持つことを明示的に表現できる。このことは、日本語における一部の格交替を合理的に説明する。[2] このようなことから、汎化された意味機能と格助詞などの統語情報を結びつけるための手がかりを与える情報を含んだ言語資源の枠組みは重要であると考える。

[私が/から $_i$]　[それを $_j$]　[彼に $_k$]　伝え ます。

伝える.v

$$\left[\begin{array}{l}\text{cause(a\ \ ect}(i,j)\text{, go}(j,\ \left[\begin{array}{l}\text{from(locate(in}(i)))\\ \text{to(locate(at}(k)))\end{array}\right]))\\ \\ \text{for}\left[\text{cause(a\ \ ect}(k,j)\text{, go}(j,\ \left[\text{to(locate(in}(k)))\right]))\right]\end{array}\right]$$

図 1: 述語「伝える」の LCS 構造と格交替 [1]

プロジェクトの初期段階として、我々は既存の LCS 理論を、実際のテキストに対する述語項構造アノテーションに適した形に拡張しながら、LCS 構造に基づい

[1] 本稿で LCS 構造を記述する際は、説明の簡単のため、各基本述語が取る細かな属性値は省略してある。

[2] 例えば、図 1 の述語「伝える」では、項 $i$ が、*affect* の最初の項と *from* の内側の項として、二回出現する。 我々は、このような LCS 構造を持つ多くの動詞について、ガ格とカラ格の交替が出来ることを確認している。

[私は $_i$]　[彼から $_k$]　[招待を $_j$]　受け た。

受ける.v

$$\left[\begin{array}{l}\text{cause(a ect}(i,j)\text{, go}(j, \left[\text{to(locate(in}(i))\right])) \\ \text{comb}\left[\text{cause(a ect}(k,j)\text{, go}(j, \left[\begin{array}{l}\text{from(locate(in(k)))} \\ \text{to(locate(at}(i))\end{array}\right])\right)\right]\end{array}\right]$$

図 2: 述語「受ける」の LCS 構造

た述語項構造アノテーションの枠組みを構築した。さらに６０個の日本語動詞に対する LCS 辞書を作成し、このうち３０の動詞については、アノテーションのパイロット作業を行った。本稿では、我々がコーパス作成のために用いた枠組みを紹介すると共に、日本語動詞の LCS 辞書作成作業について、現在の状況を報告する。

## 2　枠組み

PropBank や FrameNet といった述語項構造がタグ付けされた既存の言語資源と同様に、我々の作成するコーパスも、（１）各述語の項構造（LCS 構造）を語彙項目とし、各項についての意味情報を含む、フレーム辞書（LCS 辞書）と、（２）辞書情報に基づいてタグ付けされた文書集合、の二つの資源からコーパスを構成する枠組みを取る。また、本プロジェクトを通じでタグ付けする情報は、日本語の代表的な係り受け構造解析済みコーパスである、京都大学テキストコーパス [7] と同一の文書上に、その上位レイヤーとして情報を付与することとする。これらのアプローチを採用する根拠は、それぞれの述語について、その統語的な項の完全なリストを与えることと、項の統語構造と意味構造の間の関係を捉えることが肝心であると考えるところにある。

PropBank と FrameNet では、語彙項目は各述語の異なる語義または概念であり、語彙項目には、その語義や概念に特有の役割として定義された、項の意味ラベル情報を含んでいる。我々の場合、語彙項目は各述語に対する複数の異なる LCS 構造である。我々の LCS 構造は、基本的に Jackendoff の LCS 理論に従っているが、実世界の文書に出現する多様な種類の述語に対応出来るよう、理論面での被覆率を向上させるために、幾つかの理論的拡張を施している。最も大きな変更は、新しい基本述語 *combination* の導入で、これは、一つの述語の中に存在する複数の出来事を記述出来るようにするために導入したものである。この変更は、Jackendoff が提案する基本述語 *EXCH* の自然な一般化であるが、実際のデータに出現する様々な述語に対して LCS 構造を記述する上では、本質的に重要である。具体的には、京都大学テキストコーパスに頻出する６０の日本語述語（動詞、及びイベント性の名詞）に対して我々が作成した１０９の LCS 構造のうち、４１の構造 (37.6%) では、複数のイベントを含んでいた。それだけでなく、このうち幾つかの述語に関しては、複数のイベント記述なしに、（イベントの一部が欠ける形であっても）正しい意味構造を表現することが困難であった。例えば、図 2 の述語「受ける」は三つの統語的な項 *i, j, k* を含んでいるが、二つある式のうち、最初の式だけを用いて述語の意味を表現しようとしても、項 *k* については、正しい解釈のもとに最初の式に埋め込むことが出来ない。

我々の枠組みでは、意味役割は大きく二つのクラスに分類出来る。一つは、その項の意味が、LCS の基本述語を通して規定される役割のクラスであり、我々はこれらを主要役割と呼ぶ。主要役割については、タグ付与者は、従来の述語項構造コーパスでいうところの意味タグの代わりに、文書中のタグ付与対象となる述語の項に対して、LCS 構造中の表記に従って、項の ID（例えば、*i, j, k* など）を付与していく。既に述べた通り、LCS の基本述語が持つそれぞれの項のスロットは、おおまかに従来の主題役割と対応しているため、英語圏での既存の資源に似た、しかし、項同士の関係がより構造化された意味役割の情報が、LCS の構造を通じて各項に付与されることになる。

もう一つのクラスは、一般的に多くの述語と共に表れることができ、LCS の基本述語の意味として表現されない役割を含む。我々はこれらを周辺的役割と呼び、これらには、例えば、 *time*, *place*, *manner* などが含まれる。このクラスに含まれる意味役割を定義するため、我々はまず、現代日本語文法 [1] に記述された５２種類の格助詞の用法を参考にし、そこから LCS の基本述語で表現される意味機能に相当するものを排除した。次に、タグ付与時の曖昧性排除を考慮し、類似するカテゴリのものを統一し、最後に、京都大学テキストコーパスを事前に解析していた上で観測された新たな意味役割、格助詞以外の表現による用法などを追加し、最終的に２１の役割を含む集合とした。

図 3 に、我々のアノテーションプロセスの概要を示す。タグ付けの最中、タグ付け作業者は、まず、タグ付けの対象とする述語を見つけ、次に LCS 辞書の中にある対象述語の辞書項目の中から、正しい LCS 構造を選ぶ。次に、選び出した LCS 構造と、周辺的役割のリストを見ながら、上記の二つのタイプのラベルを文内の項についてタグ付けしていく。本プロジェクトで

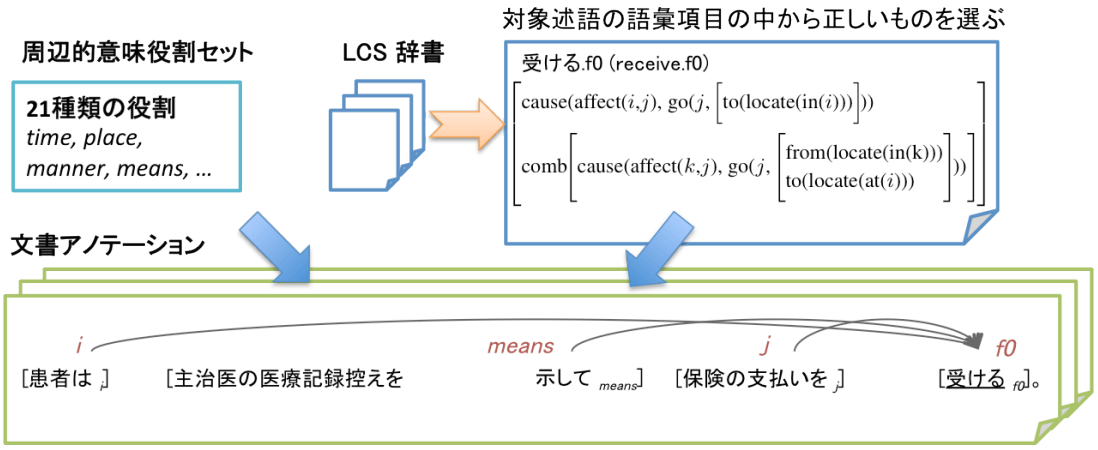


図 3: コーパスアノテーションの枠組み

は、我々は基本として、統語構造と意味構造の繋がりをまず明らかにするという立場をとっているため、統語的に何らかの結びつきのある項を対象にタグを付与するという方針でタグ付けを行なっている。したがって、現状では、我々のタグ付け作業にはゼロ照応や共参照関係は含まれていない。

我々のプロジェクトは、現在初期段階にある。現在までに、事前に作成した LCS 辞書に含まれる３０の述語について、各述語毎に約１００事例のタグ付けを行い、初期のタグ付けガイドライン策定を行ったところである。最終的なプロジェクトの目標は、京都大学テキストコーパスに含まれる全ての文に対し、述語項構造をタグ付けし、公開することである。

## 3 LCS 辞書の構築

初期のタグ付け作業を開始する前に、我々は京都大学テキストコーパスに頻出する６０の述語について、LCS 構造を記述した。ここでの目的は、現実世界のデータに出現する様々な種類の動詞を被覆できるよう、LCS 理論を拡張し、各述語に対して、一貫性のある LCS 構造を作成する方法を確立することであった。

異なる述語間の LCS 構造の一貫性を維持するため、我々は次の三つの方法を取った。第一に、LCS 構造は、著者のうち一名が京都大学テキストコーパスにおける対象述語の事例を見ながら、全て一人で記述した。この際、語義と格フレームの被覆率を向上させるため、インターネットで利用可能な辞書であるデジタル大辞泉、[3]及び、大規模な Web コーパスより自動獲得した格フレーム辞書である京都大学格フレーム [4] の情報を参考にした。第二に、辞書作成と平行して、対象述語において最も焦点の当てられた主たるイベント或いは状態を表す、LCS 構造の第一番目の式を決めるための決定木を開発した。これは、我々が、LCS 構造の最初の式について、有限個のスケルトンフレームを作り、決定木を用いてその内の一つを選択する手法を取ったということである。第三に、我々は、二つの述語の間の語彙的な含意関係を人手で確認し、これらの述語の構造が類似するように LCS 構造を構築した。より具体的には、我々は、LCS 構造上に、含意関係を保持するような複数の書き換え規則を定義し、これらのルールを構造間の整合性を確認するためのチェック機構として利用した。直感的には、これらの規則は、FrameNet で定義されるフレーム間の意味的なグラフ構造に似た階層構造を LCS 構造の間に構築する。我々の規則によって実際に構築される階層構造の一例を、図 4 に示す。[4]

LCS 理論に関する幾つかの拡張の後、我々は、追加の修正なしに６０動詞に対する１０９のフレームを作成することに成功した。この結果より、我々の拡張した LCS 理論は少なからず安定したものであると期待しているが、一方で、幾つかの動詞については、LCS 理論のさらなる拡張が必要となることも発見している。例えば、「つながる」や「統一」のような、相互交換に関する動詞に見られる格交替の現象（英語では、Levin の交替クラス 2.5 [8] に相当）や、これらの動詞の意味表現は、それぞれの項について、対象となる実体の数を考慮する事なしに合理的な説明を与えることは困難である。これらの動詞も含め、今後もより多く

[3]http://dictionary.goo.ne.jp/jn/より利用可能

[4]説明のため、図には我々が京都大学テキストコーパスに頻出の述語として取り出した 60 動詞に含まれない動詞が存在する。

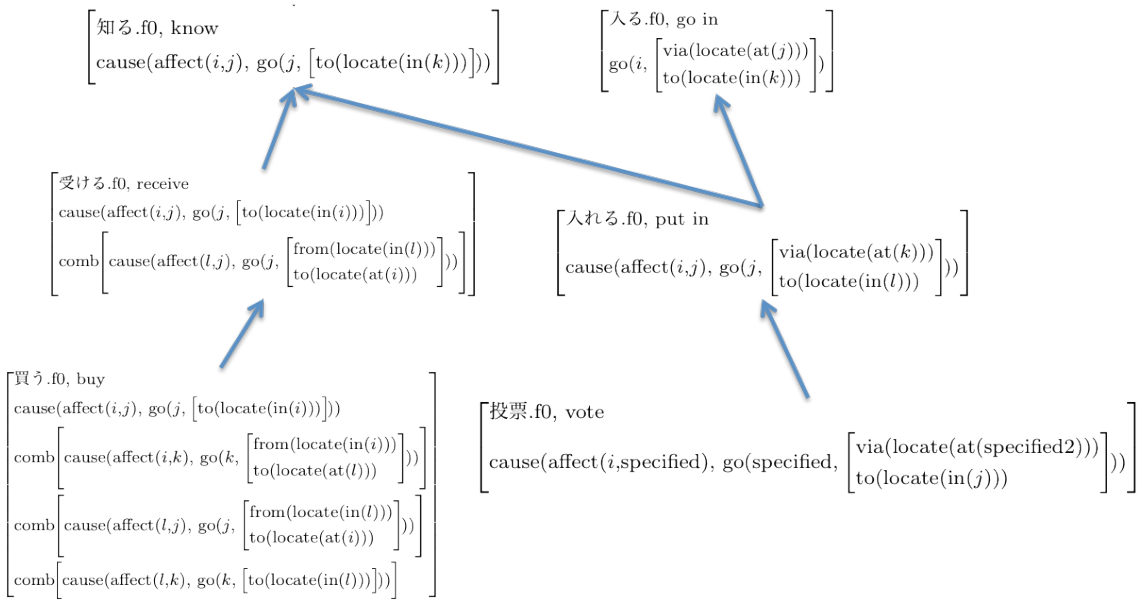


図 4: LCS 構造間に定義した書き換え規則によって構築された LCS の階層関係

の述語を被覆できるよう、理論の拡張と LCS 辞書の構築を続けていく予定である。

## 4 結論

本稿では、我々が現在進めている、より意味的な側面に動機づけされた日本語の述語項構造タグ付きコーパスの開発について、コーパス設計の枠組みと、これまでの開発で得た知見を報告した。我々は、述語項構造を有限個の基本述語によって意味の構造として分解して記述する LCS 理論を利用して、各項の持つ意味の属性をより明確にしたタグ付けフレームワークを提案した。また、Jackendoff の LCS 理論を拡張し、表現の一般性を向上させることで述語の被覆率を向上させ、京都大学テキストコーパスに頻出の６０の日本語述語に対して LCS 構造を記述することに成功した。タグ付け作業としては、３０述語に対して各１００事例のパイロットアノテーションを完了し、この過程で、逐次的にタグ付けガイドラインを策定したところである。

今回、我々は開発の過程において、意味的に関係の深い述語の間の LCS 構造に一貫性を保つ努力を行なってきたが、一方で、この一貫性の評価に関しては、将来的に、定量的、定性的な評価が必要不可欠である。このため、今後は、前節で示したような LCS 構造間における含意関係の計算手法を定式化し、FrameNet の持つ Frame 間の関係グラフ等との比較を行なっていく予定である。

## 参考文献

[1] Study group of Japanese descriptive grammar, editor. *Contemporary Japanese Grammar*, Vol. 2. Kuroshio Press, 2009.

[2] Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. Annotating a japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, pp. 132–139. Association for Computational Linguistics, 2007.

[3] Ray Jackendoff. *Semantic Structures*. The MIT Press, 1990.

[4] D. Kawahara and S. Kurohashi. Case frame compilation from the web using high-performance computing. In *Proceedings of LREC-2006*, pp. 1344–1347, 2006.

[5] Paul Kingsbury and Martha Palmer. From Treebank to PropBank. In *Proceedings of LREC-2002*, pp. 1989–1993, 2002.

[6] Karin Kipper, Hoa Trang Dang, and Martha Palmer. Class-based construction of a verb lexicon. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 691–696. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2000.

[7] Sadao Kurohashi and Makoto Nagao. Kyoto university text corpus project. *Proceedings of the Annual Conference of JSAI*, Vol. 11, pp. 58–61, 1997.

[8] Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press, 1993.

[9] J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson, and J. Scheffczyk. FrameNet II: Extended Theory and Practice. *Berkeley FrameNet Release*, Vol. 1, , 2006.