

BCCWJ への日本語フレームネットの意味アノテーション

小原 京子 (慶應義塾大学理工学部)[†]

Semantic Annotation of BCCWJ in Japanese FrameNet

Kyoko Hirose Ohara (Faculty of Science and Technology, Keio University)

1. はじめに

本稿では日本語フレームネット・プロジェクトにおける「現代日本語書き言葉均衡コーパス」(BCCWJ)への意味アノテーション作業について報告する (<http://jfn.st.hc.keio.ac.jp/>)。日本語フレームネット・プロジェクトでは、語彙アノテーションと全文テキストアノテーションという、二つのモードでフレーム意味論に基づく意味アノテーションを行ってきた (Ohara and Sato 2010)。これまで語彙アノテーションでは BCCWJ モニター公開データ 2008 年度版を、全文テキストアノテーションでは BCCWJ コアデータを対象としてきている。

日本語フレームネット・プロジェクトとは、フレーム意味論の枠組みでコーパスデータに意味タグ付けを行い、その結果を電子言語資源として構築するものである。アメリカ・バークレーで構築中の英語語彙情報資源 FrameNet (FN) との共同研究プロジェクトである (<http://framenet.icsi.berkeley.edu/>, cf. Hasegawa et al. 2010)。FN との互換性を保ちながら、英語との対比に着目して意味タグ付けを行っている。フレーム意味論に基づく意味タグ付けとは、ある語を用いる際に話者が想起する背景知識 (意味フレーム) とその意味的構成素である粒度の高い意味役割 (フレーム要素) を明らかにすることでその語彙項目の意味や用法を定義し、例文中の語句にフレーム要素を意味タグとして付与していく、というものである (小原 2009)。意味タグとして付与される情報が古典的な格フレームに基づく抽象的な役割関係 (Agent, Instrument, Object など) ではなく意味フレームに即して定義される精度の高いフレーム要素であることと、フレーム間関係を参照することで語彙素同士の関連付けがなされている点が特徴である (cf. Fillmore 1987)。

以下では、BCCWJ への日本語フレームネットの意味アノテーションについて、特に英語と日本語の語彙分析に必要な意味フレームの比較の観点から述べる。第 2 節では語彙アノテーション、第 3 節では全文テキストアノテーションについて報告し、第 4 節で今後の方向性について述べる。

2. 日本語フレームネットの語彙アノテーション

語彙アノテーションとは、語彙項目 (lexical unit, LU) ごとに BCCWJ の中からアノテーション対象例文を選び、タグ付けしていくモードである (小原 2008a)。具体的には以下の四つの作業を行っている: 1) 動詞・形容詞・副詞・事態性名詞の、BCCWJ における出現頻度の高い語彙素 (lexeme) から順にアノテーション対象を決める; 2) その語彙素が喚起する意味フレーム (言語の発話や理解の際に必要な、体系的知識構造) を同定する。実際には、FN 上で英語語彙分析のために既に定義された意味フレームの中から当該日本語語彙素に適用できるものを探す。FN 上の既成の意味フレームの中に適切なものがない場合には、i) 英語語彙分析に必要なにもかかわらず、FN 上で今現在まだ定義されていないだけなのか、あるいは ii) 英語語彙分析には不要なために FN 上で定義されていないが日本語

[†] ohara@hc.cc.keio.ac.jp

語彙の意味分析のためには新たに定義が必要な意味フレームなのか、を検討する；3) コーダンサーを用いて BCCWJ から当該語彙素を検索し、さらにその中から当該意味フレームに関与する例文のみを選別する。さらに、共起語や結合価パターンを考慮しつつ、アノテーション対象とする例文を選定・抽出する；4) 抽出した文に、アノテーションツールを用いてフレーム要素¹、統語情報、文法情報などのタグを付与する、の四つのプロセスである。

日本語フレームネット・プロジェクトでは、英語語彙分析のために FN で定義された意味フレームが類型論的に異なる日本語の語彙意味記述にどこまで適しているのかを検証している。従って、語彙アノテーション作業においても、日本語固有の意味フレームやフレーム要素を定義する必要があるのか、必要があるとしたらどのような意味フレームとフレーム要素が必要なのかを検討してきた。次節で見るように日本語フレームネットで意味フレームを独自に定義する必要に迫られたケースはこれまでにさほど多くはなかった。むしろ、日英両言語の語彙に関与する意味フレームの違いとは、個々の意味フレームのレベルではなく、体系的な違い、すなわち意味フレームの立て方に関する違いであることがわかってきた。

たとえば、(1a) のように「散る」という語に関与する意味フレームを同定するケースを考えてみる。まず、この自動詞「散る」に対応する他動詞「散らす」に関与する意味フレームとしては、scatter という英語の対応語を手掛かりに Dispersal フレーム（「動作主が起点(狭い空間)から終点(広い空間)へと個体の集合体をばらまく」)が同定できる(1b)。しかしながら、既存の意味フレームのうち自動詞「散る」に関与していると考えられるのは物の移動に関する一般的なフレーム Motion フレーム（「個体が起点を出発し、経路を通ることにより終点に行く」)のみである(1a)。

- (1) a. 桜の花びらが散る Motion。
b. 桜の花びらを散らす Dispersal。

他の自動詞「刑死する」とそれに対応する他動詞「処刑する」の場合も同様である(2)。まず、他動詞「処刑する」に関与する意味フレームとしては、execute という英語の対応語を手掛かりに Execution フレーム（「処刑者がある行為を理由にある個人に罰として死を与える」)が同定できる(2b)。ところが、既存の意味フレームのうち自動詞「刑死する」に関与していると考えられるのは、死に関する最も一般的なフレーム Death フレーム（「ある参加者の死についての描写」)のみである(2a)。

- (2) a. 死刑囚が刑死する Death。
b. 死刑囚を処刑する Execution。

¹ フレーム意味論ならびにフレームネットにおけるフレーム要素とは、それぞれの意味フレームに依存した具体的な意味的要素である。これに対し、VerbNet (<http://verbs.colorado.edu/~mpalmer/projects/verbnet.html>) では汎用的な意味役割を付与している。VerbNet の汎用的意味役割を用いて FrameNet と PropBank (<http://verbs.colorado.edu/~mpalmer/projects/ace.html>) の「意味役割」の汎化を行った研究に松井他(2010)がある。汎用的意味役割の課題については、上代日本語コーパスへの意味アノテーションに基づくホーン(2011)を参照されたい。

b. 死刑囚を処刑する Execution

英語の語彙意味分析のために定義された既存の意味フレームでは、動詞の自他交替に関しては他動詞的観点から定義された意味フレームが比較的多く、自動詞的観点から定義されたものは少ない。上で見た Dispersal フレームと Execution フレームも他動詞的観点から定義されたものである。動詞の自他交替に関して、他動詞的観点と自動詞的観点のそれぞれから意味フレームが定義されているケースはほとんどなく、例外は Detaching フレーム（「動作主がある物を他の物から外す情景」と Becoming_detached フレーム（「ある物が他の物から外れる情景」）の対くらいである²。Filling フレーム（「容器をある物でいっぱいにする」と Fullness フレーム（「容器がある内容物でいっぱいになっている状態」）も一見それぞれ他動詞的観点と自動詞的観点から定義された意味フレームの対のように見えるが、実際にはそうではない。Filling フレームが他動詞的観点から定義されているのに対し、Fullness フレームの方は自動詞的観点から定義されているのに加えて状態相についての意味フレームでもある。したがって、Filling フレームは他動詞「満たす」に関連していると考えられるが、Fullness フレームは自動詞「満ちる」に関与する意味フレームではない。「満ちる」の語彙的アスペクト（動作相、Aktionsart）は開始相であり、状態相ではないからである³。日本語動詞は一般に語彙的アスペクトが開始相のものが多い。

このように既存の意味フレームは当初英語の語彙意味分析のために定義されたものなので、日本語の語彙意味分析に用いようとする、定義の一部に含まれ前提とされている視点（他動詞的・自動詞的）や語彙的アスペクト（状態相・開始相）が日本語の語彙のそれらとずれていることが多々あることがわかった。今後の語彙アノテーション作業、特にフレーム同定作業ならびにフレーム定義作業においても十分この点を考慮する必要がある。

3. 日本語フレームネットの全文テキストアノテーション

全文テキストアノテーションでは、テキスト内のすべての文の、意味フレームを喚起 (evoked) するすべての語彙項目に対してアノテーションを行う。BCCWJ コアデータ書籍ジャンル（総数 84 ファイル）の固有表現以外の語彙項目を対象とした。また、全文テキストアノテーション結果を基に、FN で既に定義された意味フレームがどこまで日本語の語彙記述にも有効であったかを調べた (cf. 小原 2011)。

全文テキストアノテーションを BCCWJ コアデータのサンプルごとに施すことのメリットとしては以下が挙げられる。まず、フレーム意味論に基づく意味タグ付きコーパスが作成できる。また、BCCWJ のサンプルごとに、意味フレーム（すなわち語義）の分布や、結合価パターン、ゼロ代名詞の分布などを詳細に調べることができる。将来的には BCCWJ コアデータに対する他の体系に基づくアノテーションと比較・統合することも視野に入れている。

各語彙項目の喚起する意味フレームを同定する際の手順は以下の通りである。まず、FN

² ちなみに、意味フレーム名の冒頭に “Becoming_” と付くものは自動詞的観点から定義された意味フレームであるが、2011 年 7 月 10 日現在 1056 ある意味フレームのうち、Becoming_a_member フレーム、Becoming_aware フレーム、Becoming_detached フレーム、Becoming_dry フレーム、Becoming_separated フレームの 5 フレームのみである。

³ 「外す」関連では、前述の Detaching フレーム（他動詞的観点）と Becoming_detached フレーム（自動詞的観点・開始相）に加えて、Being_detached フレーム（状態相）も定義されている。

上で英語語彙分析のために既に定義された意味フレームの中から、当該日本語語彙項目に適用できるものを探す。もし FN 上の既成の意味フレームの中に適切なものがない場合には、
i) 英語語彙分析に必要なだが、FN 上で今現在はまだ定義されていないだけなのか、あるいは
ii) 英語語彙分析に不要なため FN 上で定義されていないが、日本語語彙の意味分析のためには新たに定義が必要な意味フレームなのか、を検討する⁴。

このようにして、FN 上で既に定義された意味フレームのうちどの程度が BCCWJ コアデータ書籍ジャンルの語彙記述にも使うことができたかを調べた。その結果、書籍ジャンルのサンプルにおける FN 上の意味フレームの適合率は平均 81 パーセントであった。適合率の算出に当たっては、異なり語 (type) 数ではなく延べ語 (token) 数を用いた。

適切な意味フレームが FN で既に定義された意味フレームの中に見つからず、意味フレーム名が付与できなかった語彙項目を品詞別にまとめたのが(3)である。括弧内の数字は 179 語の異なり語のうちその品詞に分類されるものの数を示す。

(3) 意味フレーム名が付与できなかった語彙項目 (品詞別) (括弧内は異なり語数)

a. 形容詞 (1 語)

あらい

b. 接続詞 (4 語)

だから、しかし、ならば、すなわち

c. 形状詞 (13 語)

好意的、当然、一方的、文字通り、圧倒的、地理的、分野的、のろま、順調、簡単、凄絶、徹底的、科学的

d. 動詞・動詞句 (12 語)

悪びれる、過ごす、占める、建ち並ぶ、向きあう、する、潜む、遊ぶ、さし出す、向く、間に合う、気をつける

e. 副詞 (21 語)

実際のところ、もちろん、必ずしも、一切、しっかり、ギリギリ、一般に、たとえば、半ば、最も、実際、言い換えれば、はじめて、例えば、多分、単に、少なくとも、たまたま、つかつか、代々、もともと

f. 事態性名詞 (22 語)

営業、制御、通行、参照、捨象、仲介、紹介、関連、生活、転勤、出血、観閲、闊遊び、体験、経験、解剖、出版、棚割、お使い、刑死、牢死、埋葬

g. 名詞 (106 語)

支店、肉声、セールス、手づくり、基準、盤、常識、神、精神、精神構造、神霊、海魚、借り、玩具、行き、モノ、聖書、単位、一方、核、スケール、印税、こと、社会、民間、産業主義、かけがえ、ホームレス、毛布、巢、諺、育ち、魚、白豚、国語、紙、体質、良性、悪性、細胞、割合、組織、元気、ぬいぐるみ、犬、ペット、ワン、畳、

⁴ ドイツの SALSALSA プロジェクトにおいても同様の方針がとられている (<http://www.coli.uni-saarland.de/projects/salsa/page.php?id=index>)。

ソファ、抱っこ、声、心、音楽、闇、画面、モデル、自然、事象、法則、原理、現象、遊び、弾力性、バネ、授業、内容、呪力、レーダー、先、共通、点、民俗、文化、港、岸壁、博士、一説、自分、例、あて字、死体、神仏、内臓、川獺、罪人、前例、構成、ユニット、障子、週別、襖紙、同週、ポスト、門、通い、暮らし、志士、墓、夢、屍骸、土、野犬、墓地、侠客、情況、筋肉

意味フレーム名が付与できなかったこれらの語彙項目のほとんどは上記の i) の「英語語彙分析のためにも必要と考えられるが、FN 上で今現在はまだ定義されていない」ケースと考えられ、ii) の「英語語彙分析に不要なため FN では定義されていないが、日本語語彙の意味分析には新たに定義が必要と考えられる意味フレーム」はごく少数にとどまった。すなわち、異なり語 179 語のうち、日本語フレームネットで独自に意味フレームを定義する必要があると考えられた語彙項目は、(3)で下線を付けた 4 語（「畳」、「障子」、「襖紙」、「侠客」の名詞）のみであった。

i) の、「英語語彙分析にも必要と考えられるが、FN 上で今現在はまだ定義されていない」ため意味フレームが付与できなかった語彙項目の中には、接続詞(3b)や副詞(3e)が含まれている。これは、FN では副詞や接続詞のアノテーションがまださほど進んでいないことが原因と考えられる。さらに、副詞(3e)の中には、「実際のところ」、「実際」、「もちろん」、「言い換えれば」などのメタテキスト的な文副詞が多く含まれている。これらの語彙項目の意味や用法は文の談話的状況と密接に関係しているので、これらが喚起する意味フレームやそのフレーム要素を意味論的に定義するのは容易ではないであろうと予測できる。一方、意味フレームが付与できなかった形容動詞(3c)、動詞・動詞句(3d)、事態性名詞(3f)、名詞(3g)が関与する意味分野は多岐にわたることがわかった。

また、複数アノテータが付与した意味フレーム名のうち、どの程度が一致しているかを調べた。全文テキストアノテーション作業では、第一段階として、技術翻訳専門の翻訳者にアノテーション対象である日本語語句の英訳を考えてもらった。さらに、FN データベースを検索してその英訳語が喚起する意味フレームを調べてもらった。元の日本語語句もその英訳語が喚起する意味フレームと同じ意味フレームを喚起すると考え、日本語語句の喚起する意味フレームを同定してもらった。第二段階では、語彙アノテーション作業経験が 1 年以上のアノテータに第一段階の意味フレーム同定結果を検討してもらった。さらに第三段階では、筆者が最終的な意味フレームの同定を行った。その結果、第二段階と第三段階とでは一致率が平均 66 パーセントであった。アノテータ間の一致率が比較的低いことは、意味フレーム名の付与作業がかなり高度な作業であることを示唆している。また、意味フレーム同定の際に、その語句の英訳語を決め、その英訳語の喚起する意味フレームを同定する、と言う作業が介在することも関係していると考えられる。

4. おわりに

BCCWJ への日本語フレームネットの意味アノテーションについて報告した。特に、FN 上で定義済みの意味フレームの日本語語彙意味記述への適用性について、語彙アノテーションと全文テキストアノテーションの両方から考察した。今後も、英語の語彙意味記述には必要ないが日本語には必要な意味フレームとはどのようなものか引き続き検討していく必要がある。

付記

本稿は、『特定領域「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集』に掲載した小原・斎藤・藤井・佐藤（2011）と小原・加藤・斎藤（2011）を基に執筆した。本稿で報告したアノテーション作業にあたりご協力をいただいた木越壽子氏、前木香織氏、アレクサンドル・カバチュ氏、李陽氏、加藤淳也氏、萩野弘也氏、黒川奈桜子氏に御礼申し上げます。

主要文献

- 小原京子(2009) 「5.2.6 格文法」、「5.2.7 フレーム意味論」、『言語処理学事典』pp.270-275, 364-371. 共立出版.
- 小原京子(2011) 「日本語フレームネットの全文テキストアノテーション：BCCWJ への意味フレーム付与の試み」、言語処理学会第 17 回年次大会予稿集.
- 小原京子、斎藤博昭、佐藤弘明、藤井聖子 (2011). 「研究活動・成果の総括：日本語フレームネット班 BCCJW と意味フレームに基づく語彙・構文複合資源の構築」、『特定領域研究「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集』, pp. 231-240.
- 小原京子、加藤淳也、斎藤博昭 (2011). 「日本語フレームネットにおける BCCWJ への意味アノテーション」、『特定領域研究「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集』, pp. 513-518.
- ホーン、スティーブン・ライト(2011). 「上代日本語のコーパスにおける意味役割の付与」国立国語研究所. 2011 年 1 月 20 日.
- 松林優一郎、岡崎直観、辻井潤一 (2010). 「自動意味役割付与における意味役割の汎化」『自然言語処理』pp.59-89. Vol.17, No. 4.
- Fillmore, Charles J. (1987). “A private history of the concept ‘frame’.” Dirven, Rene and Radden, Gunter. (Eds.), *Concept of Case*. Gunter Narr Verlag, Tübingen. pp.28-36.
- Fillmore, Charles J. and Collin Baker (2010). “A frames approach to semantic analysis.” In Heine, Bernd and Heiko Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis*. pp.313-339. Oxford University Press.
- Hasegawa, Yoko, Russell Lee-Goldman, Kyoko Hirose Ohara, Seiko Fujii, and Charles J. Fillmore (2010). “On expressing measurement and comparison in English and Japanese.” In Boas, Hans C. (Ed.), *Contrastive Studies in Construction Grammar*. pp.169–200. Amsterdam: John Benjamins Publishing.
- Ohara, Kyoko Hirose and Hiroaki Sato (2010). “Investigating Japanese FrameNet Data with FrameSQL.” Sixth International Conference on Construction Grammar (ICCG-6). Charles University, Prague, Czech Republic. September 5th, 2010.

関連 URL

日本語フレームネットホームページ : <http://jfn.st.hc.keio.ac.jp/ja/index.html>
FrameNet ホームページ : <http://framenet.icsi.berkeley.edu/>