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への意味フレーム名付与の試み

小原 京子（慶應義塾大学）

1. はじめに

本論文では、日本語フレームネット（略称 JFN）・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)への意味フレーム名の付与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では、BCCWJモニター公開データを対象に、テキスト内に出現する自立語すべてへの意味フレーム名の付与（全文テキストアノテーション）を試みている（http://www.tokuteicorpus.jp/）。本論文ではBCCWJの「書籍」ジャンルのテキストへの作業経過について報告する。特に、1) 英語フレームネット[1]（略称 FN）上の意味フレーム定義の適合率、2) 日本語固有の意味フレーム定義の必要性、3) アノテータ間の意味フレーム名付与の一致率について述べる。 英語・日本語フレームネットの枠組みに基づく意味フレーム名付与済みコーパスは、意味タグ付きコーパスとして情報検索・テキスト要約などの自然言語処理アプリケーションに利用されることが期待される。

フレームネット・プロジェクトでは、フレーム意味論とコーパスデータに基づき英語のオンライン語彙情報資源を構築中である（http://framenet.icsi.berkeley.edu/, Fillmore & Baker 2010）。日本語フレームネットは2002年から始まった日本語語彙情報資源構築プロジェクトで、フレームネットとの連携のもとに進められている（Ohara & Sato 2010, Tagami et al. 2009, cf. Hasegawa et al. 2010）。フレームネットの手法で、コーパスデータを用いて語の意味・用法の分析を行い、オンライン日本語語彙情報資源の雛型を構築している。英語語彙分析のためにフレームネット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討するのが主な目的の一つである。

本論文の構成は以下のとおりである。まず、次節で日本語フレームネットにおける全文テキストアノテーション、すなわちBCCWJへの意味フレーム名付与作業の概要について述べた後、第3節では英語語彙の意味分析のために定義された意味フレームがどこまで日本語テキストのアノテーションに適用できたかを、適合率の観点から報告する。それを踏まえ、第4節では日本語固有の意味フレームとして新たに日本語フレームネット上で定義が必要な意味フレームについて考察する。第5節ではアノテータ間の意味フレーム名付与作業の一致率について述べる。

2. 日本語フレームネットの全文テキストアノテーションとBCCWJ

日本語フレームネットでは語彙項目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BCCWJへのタグ付けを行ってきた。語彙項目アノテーションとは、語彙項目ごとにBCCWJの中からアノテーション対象とする例文を選びタグ付けしていくモードである。これに対して全文テキストアノテーションとは、特定のサンプルテキスト内の全ての文の、意味フレーム（言語の発話や理解の際に必要となる、体系的知識構造）を喚起（evoke）する全ての語彙項目に対してタグ付けしていくモードを指す。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（人手で形態素解析結果を修正した、各ジャンルのサンプルのサブセット）を対象に分析・アノテーションを行ってきた。

全文テキストアノテーションとは、テキスト内のすべての文の、意味フレームを喚起するすべての語彙項目に対してアノテーションを行うことである。固有名詞以外の語彙項目が対象である。本論文では、BCCWJコアデータ書籍ジャンルの各サンプル（総数84ファイル）の冒頭10文の意味フレーム喚起語への意味フレーム名付与結果について論じる。

全文テキストアノテーションは、語彙アノテーション同様にJFNDesktopというアノテーションツールを用いて行っている。全文テキストアノテーション結果表示用ツールは、語彙アノテーション結果表示用ツールとは別に開発した。

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

[1] 正式名称はFrameNetであるが、本論文では日本語フレームネットと比較して議論する際に必要に応じてFrameNetを「フレームネット」ではなく「英語フレームネット」と表記することにする。オンライン語彙資源構築にフレームネット同様の枠組み・手法を用い、フレームネットと共同研究を行っているプロジェクトとしては、日本語フレームネットの他に、スペイン語フレームネット(http://gemini.uab.es:9080/SFNsite)やドイツ語フレームネット(http://gframenet.gmc.utexas.edu/)がある。

ーションと比較・統合することも可能となる。

3. 英語フレームネット上の意味フレームの適合率

日本語フレームネットでは、まず英語フレームネットの英語語彙分析のための意味フレーム定義が日本語語彙分析にも適用できるかを検討し、英語フレームネット上に適切な意味フレームが存在しない場合には、i) 英語フレームネット上でたまたま未定義なだけなのか、ii) 英語の語彙分析には不要だが日本語語彙の意味分析には必要な意味フレームなのか、を考察する。

この方針を全文テキストアノテーションにも適用し、英語フレームネット上の意味フレームがどの程度BCCWJコアデータ書籍ジャンル上の語彙記述に用いることができるかを調べた。その結果、書籍ジャンルのサンプルにおける英語フレームネットの意味フレームの適合率は平均 82 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）ではなく延べ語（token）を用いた。

BCCWJ コアデータ書籍ジャンルのサンプルにはフィクションとノンフィクションの両方が含まれるが、概してノンフィクションの方がフィクションより適合率が高かった。ノンフィクションで平均 81 パーセントであったのに対し、フィクションでは平均 90 パーセントであった。

4. 日本語固有の意味フレーム

上の第 3 節でみたように、サンプル上に出現する日本語の語彙項目の意味を表すのに適切な意味フレームが英語フレームネット上に見つからなかった場合、i) 英語の語彙分析にも必要だが英語フレームネット上でまだ定義されていないだけなのか、ii) 英語の語彙分析には不要だが日本語語彙の意味分析には必要な意味フレームなのか、を検討した。その結果、適切な意味フレームが英語フレームネット上で見つからないケースのほとんどは i) であり、ii) は稀であることがわかった。すなわち、異なり語 40 語のうち、ii) に該当するのは 1 語（「神霊」）のみにとどまった。i) の中には、「実際のところ」、「もちろん」、「もっとも」などの文副詞、「だから」、「しかし」、「ならば」などの接続詞が含まれていた。英語フレームネットでは副詞や接続詞のアノテーションがまだ進んでいないことが原因だと考えられる。

5. アノテータ間の意味フレーム名付与の一致率

複数アノテータが付与した意味フレーム名がどれだけ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においては、まず、第一段階として通常主に技術翻訳に従事しているプロの翻訳者にBCCWJのサンプル上の日本語語句の文脈を考慮した英訳を考えてもらい、その英語の語句を英語フレームネットデータベースで検索し元の日本語語句にふさわしい意味フレーム名を同定してもらった。第二段階では日本語フレームネットの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翻訳者によるアノテーション結果を再検討してもらった。さらに第三段階で筆者が最終的な意味フレーム名の同定を行った。その結果、第一段階と第三段階とでは意味フレーム名の一致率が平均 58 パーセント、第一段階と第三段階とでは一致率は平均 68 パーセントであった。このように、複数アノテータが付与した意味フレーム名の一致率が比較的低いことは、日本語フレームネットによる意味フレーム名付与作業がかなり高度であることを示唆している。また、意味フレーム同定に当たって英語フレームネットのデータに照らし合わせる必要があることも関係していると考えられる。

6. おわりに

以上、本論文では、日本語フレームネットにおけるBCCWJ コアデータ書籍ジャンルへの意味フレーム名の付与作業について報告した。英語フレームネット上の意味フレームの適合率については 80 パーセント以上であった。さらに、今現在までのアノテーション作業においては日本語の語彙意味分析のために固有の意味フレームを定義しなければならないケースはさほど見当たらなかった。今後も日本語固有の意味フレームとはどのようなものかについて検討していく必要がある。また、アノテータ間の意味フレーム名付与一致率を向上させるにはどうすればよいのかも考えていくべきである。

謝辞

日本語フレームネット構築には、文部科学省研究費特定領域研究「代表性を有する大規模書き言葉コーパスの構築：２１世紀の日本語研究の基盤整備」（平成１８－２２年度）による支援を受けた。

主要参考文献

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*. 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*. 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.

Tagami, H., Hizuka, H., Saito, H. Automatic Semantic Role Labeling based on Japanese FrameNet - Progress Report -, (2009). Proceedings of Conference of the Pacific Association for Computational Linguistics (PACLING2009), pp.181-186, Hokkaido, Japan, September 2009.