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

小原　京子（日本語フレームネット班班長：慶應義塾大学理工学部）†
加藤　淳也（日本語フレームネット班協力者：慶應義塾大学理工学研究科）
斎藤　博昭（日本語フレームネット班分担者：慶應義塾大学理工学部）

Full Text Annotation of BCCWJ in Japanese FrameNet

Kyoko Hirose Ohara　(Faculty of Science and Technology, Keio University)
Junya Kato　(Graduate School of Science and Technology, Keio University)
Hiroaki Saito　(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（略称 JFN）班における、「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、つまり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）を行った。本稿では、BCCWJ の「書籍」ジャンルのテキストへのアノテーション作業を中心に、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・プロジェクトは 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・プロジェク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。

本稿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 節で全文テキストアノテーション結果閲覧のためのツール、全文テキストアノテーション Web Report を紹介する。第 4 節では英語フレームネット上で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 5 節では日本語固有の意味フレ

† ohara@hc.cc.keio.ac.jp

1 正式名称は FrameNet であるが、本稿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・プロジェク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･プロジェクトの他に、スペイン語フレームネット･プロジェクト(http://gemini.uab.es:9080/SFNsite)やドイツ語フレームネット・プロジェクト(http://gframenet.gmc.utexas.edu/)がある。

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 6 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における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それらの例文に対してタグ付けするモードである。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するモードを指す。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）を対象に分析と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では、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。固有表現以外の語彙項目が対象である。本稿では、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 行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・統合することも可能となる。

３．全文テキストアノテーション Web Report

全文テキストアノテーション作業は､語彙アノテーション作業同様に JFNDesktop という、英語フレームネット用に開発されたアノテーションツールを移植・日本語化したツールを用いて行っている。図 1 は、JFNDesktop 上の全文テキストアノテーションモードでアノテーション作業を行っているところである。

アノテーション結果閲覧ツールに関しては、全文テキストアノテーション結果閲覧ツール（全文テキストアノテーション Web Report）は、語彙アノテーション結果閲覧ツール（語彙アノテーション Web Report）とは別に開発した。図 2 は全文テキストアノテーション Web Report のトップページである。BCCWJ コアデータ・テキストのうち、冒頭 10 行の全文テキストアノテーションが終了したものが表示されている。この画面でアノテーション結果を閲覧したいテキスト名をクリックすると、そのテキストのアノテーション結果が表示される（図 3)。図 3 は BCCWJ コアデータの書籍ジャンル内のテキストへの全文テキストアノテーション結果を表示したものである。青字で表示された語彙項目に対して付与された意味フレーム名の名称がその語彙項目の右下に表示されている。ちなみに意味フレーム名は英語フレームネットにおける意味フレーム名と同じものを用いており、英語で表示されている。自立語のうち青字で表示されていないものは、全文テキストアノテーション対象外（代名詞、固有表現など）のもののほか、まだ該当する意味フレームが英語フレームネットで未定義のものが含まれる。

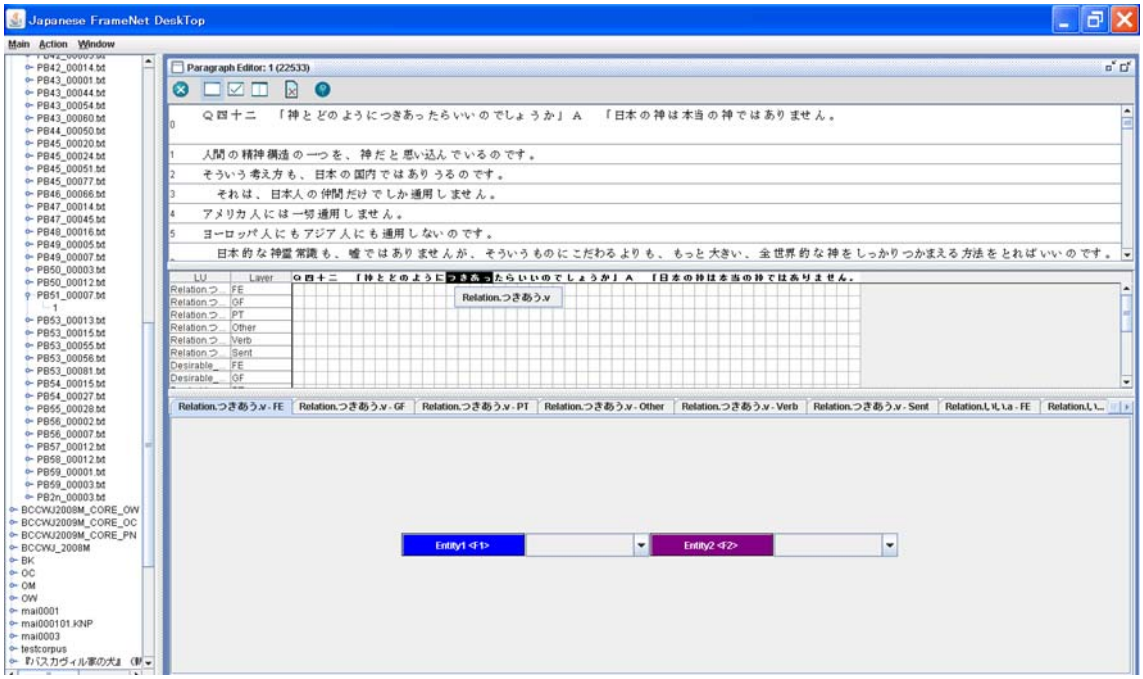

図１　JFNDesktop 上での全文テキストアノテーション作業画面

全文テキストアノテーションWeb Report_ver.2011.01.30

BCCWJ2008M_CORE_BK

- PB55_00028.txt (弁理士が答える知って得する知的財産権Q&A)
- PB40_00035.txt (宝石函)
- PB36_00006.txt (〝田舎〟社長の成功経営術)
- PB40_00003.txt (SEとして生き抜くワザ)
- PB51_00007.txt (とこしえの命を得るために)
- PB58_00012.txt (語源を楽しむ)
- PB49_00005.txt (螢の舟)
- PB11_00006.txt (ひとりの小さなおともだちが)
- PB45_00051.txt ([新編]住居論)
- PB50_00003.txt (ザ・エージェント)
- PB37_00050.txt (尼崎相撲ものがたり)
- PB43_00060.txt (企業の社会的責任)
- PB53_00015.txt (教養教育は進化する)
- PB25_00063.txt (いえづくりをしながら考えたこと。)
- PB39_00009.txt (五十メートルの戦記)
- PB43_00001.txt (授業力)
- PB54_00015.txt (医師による切らない「赤アザ・赤ら顔(浮きでた青い血管)」の最新治 療)
- PB59_00003.txt (天の前庭)
- PB26_00043.txt (犬と話ができる！)
- PB12_00001.txt (闇を歩く)
- PB33_00037.txt (子どもの感性がつくる理科授業)

図２　全文テキストアノテーション Web Report

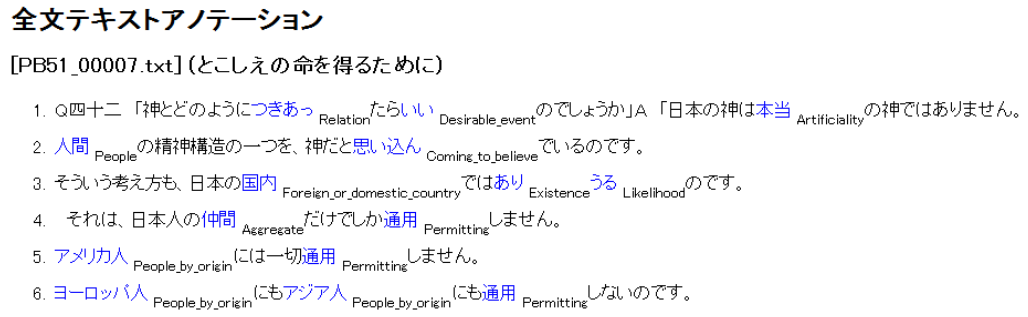

図３ 全文テキストアノテーション結果表示画面

４．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班では、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、を考察している。

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低かった。ノンフィクションで平均 81 パーセントにとどまったのに対し、フィクションでは平均 90 パーセントであった。

５．日本語固有の意味フレーム

前節でふれたように、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。異なり語 192 語のうち、ii) に該当するのは 4 語（「畳」、「障子」、「襖紙」、「侠客」）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と考えられる。

６．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 67 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。

7．おわりに

以上、本稿では日本語フレームネット班における 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平均 82 パーセントであった。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。

付記

本稿は、『言語処理学会第 17 回年次大会予稿集』に掲載した小原(2011)の一部を書き改めたものである。本稿で報告した全文テキストアノテーション作業にあたり、多大なるご協力をいただいた日本語フレームネット班研究協力者の木越壽子氏、李陽氏、並びに前木香織氏とアレクサンドル・カバッシュ氏に御礼申し上げる。

主要文献

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.

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.

Tagami, Hayato, Shinsuke Hizuka, and Hiroaki Saito (2009). “Automatic Semantic Role Labeling based on Japanese FrameNet - Progress Report -.” *Proceedings of Conference of the Pacific Association for*

Computational Linguistics (PACLING2009), Hokkaido, pp.181-186. (http://jfn.st.hc.keio.ac.jp/publications.html よりダウンロード可能)

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