日英対訳絵本の語彙から見た日本語フレームネットの評価

小原 京子[*]
ohara@hc.st.keio.ac.jp
[*]慶應義塾大学・理研 AIP

大久保佳子[**]
ookubo@jsa.co.jp
[**]日本システムアプリケーション

1. はじめに

日本語フレームネット(Japanese FrameNet, JFN)プロジェクトでは、近年 BCCWJ コーパスの書籍・白書のジャンルのコアデータに対して全文テキストアノテーションを行なってきた。具体的には、上記の各ファイルの冒頭数文に出現する自立語全てを対象に、それらの各文における意味・用法に相当する意味フレーム名（語義）を付与し、それら対象語の各項に対して「フレーム要素名・文法機能名・句タイプ名・助詞の種類」から成る注釈を付与した。BCCWJ 書籍コアデータに出現する自立語（動詞・形容詞・形状詞・名詞・副詞・接続詞）の各々の語義が既に意味フレームとして JFN 上で定義済みか（意味フレームカバー率）を調査したところ、延べ語数比で 81.3%であった（小原 2011）。しかしながら、JFN 上の意味フレームに語彙がどの程度登録済みか（語彙カバー率）については評価を行なっていなかった。[1] 本稿では、BCCWJ 書籍・白書コアデータを対象とした全文テキストアノテーション結果の評価として、日英対訳絵本に出現する自立語に関する JFN の語彙カバー率と意味フレームカバー率を調査し、今後の JFN データベース拡張に結びつけることを目指す。

2. 仮説

今回の評価では、現代日本語の基本語を対象に、JFN の語彙カバー率と意味フレームカバー率を把握することを目標とした。幼児を対象として書かれた絵本は認知的な意味で基本的な語彙のみで構成されていると仮定できる。そのため、絵本に出現する自立語に対応する意味フレームを明らかにすることで、第一言語習得という認知的な観点から基本的な意味や語彙がどのようなものであるかをある程度抽出できると考えた。具体的には、日英語版の絵本に出現する対訳文を対象に、JFN 並びに英語フレームネット（FrameNet, FN）における意味フレームカバー率と語彙カバー率を調査する。基本的な意味フレーム・語彙で不足しているものについては、本調査でもたらされる結果から新たに意味フレームを定義したり、追加語を登録したりすることで、JFN の意味フレームカバー率と語彙カバー率を上げられるはずである。今回の調査ではまず JFN における語彙カバー率と意味フレームカバー率のみを調査し、FN のそれらに関する調査は追って行うこととした。

3. 調査方法

今回使用した『英語対訳つき　とべ！アンパンマン１』は漫画形式の構成で、各コマのセリフに対応する英語対訳がついている（やなせ 1991）。ストーリーは 23 話あり、すべての会話数（コマ数に相当）は 652 であった。日本語の形態素数は延べ数で 2998 であり、1 会話文あたり約 5 形態素である。延べ語数は 676 語である。品詞別形態素数（延べ数）は表 1 のとおりである。

品詞別形態素数	
接続詞	7
感動詞	102
形状詞	24
形容詞	111
動詞	456
助詞	795
接尾辞	56
接頭辞	22

[1] フレーム意味論並びにそれを言語資源として具現化した JFN では、語がある特定の語義と対応付けられた単位を語彙項目（Lexical Unit, LU）と呼ぶ。

助動詞	372
代名詞	169
名詞	616
副詞	73
補助記号	175
連体詞	20
	2998

表１　品詞別形態素数（延べ数）

表記ごとに見ると、いくつか、「ひもじい」、「ひからびる」、「かなしばり」、「くもがくれ」などの、平易でなく頻度の低い語彙が見られ、[2]そのほかに作品特有の固有名詞類・感動詞類も通常の文章よりはやや多く含まれていることが観察された（cf. 山崎 2014, 樺島 1955, 1963）。

調査の手順は以下のとおりである。

1. 日英語版両方の絵本をテキスト化し、形態素解析にかけその結果を人手で修正する。
2. 各文中の自立語を抜き出す。
3. アノテータが各自立語に対してその語義に応じた意味フレーム名を付与する。
4. 上記３の作業結果を別のアノテータが検証する（今回は日本語版テキストのみ）。
5. 上記４を基に、日本語版テキストに出現する自立語に関して、JFN の語彙カバー率と意味フレームカバー率を調査する。

調査は用言としての用法を持つもの（形状詞・形容詞・動詞・名詞（形状詞可能）・名詞（サ変可能）・名詞（サ変形状詞可能）を対象とした。それぞれの数は表２のとおりである。

	述べ語数	異なり語数
形状詞	27	13
形容詞	111	37
動詞	456	134
形状詞可能	28	15
サ変可能	48	30
サ変形状詞可能	6	3
	676	232

表２　調査対象用言の内訳と数

出現した自立語の例を品詞ごとに(1)に記す。

(1)
形状詞：だいじょうぶ、たいへん、とうめい
形容詞：ちかい、みっともない
動詞：あそぶ、あるく、かえす、かまう
名詞（形状詞可能）：せいけつ、ひま
名詞（サ変可能）：しゅじゅつ、せんたく
名詞（サ変形状詞可能）：あんしん、じゃま

4. 結果

調査結果は表３のとおりであった。

	述べ語数	異なり語数	JFN 上の意味フレームに登録済み語（異なり語）数	該当意味フレームは存在するが未登録の語（異なり語）数
形状詞	27	13	3	7
形容詞	111	37	15	19
動詞	456	134	55	60
形状詞可能	28	15	4	9
サ変可能	48	30	11	14
サ変形状詞可能	6	3	1	2
	676	232	89	111

表３　調査結果

表３の結果から JFN 上での『英語対訳つき　とべ！アンパンマン１』日本語版に出現する用言の語彙カバー率、意味フレームカバー率を異なり語に関して求めると、表４のとおりとなった。

	述べ語数	異なり語数	語彙カバー率(%)	意味フレームカバー率(%)
形状詞	27	13	23.1	76.9
形容詞	111	37	40.5	91.9
動詞	456	134	41	85.8
形状詞可能	28	15	26.7	86.7

[2] 日本語教育語彙表（https://jreadability.net/jev/）で検索をかけたところこれら４語は該当しなかった。

サ変可能	48	30	36.7	83.3
サ変形状詞可能	6	3	33.3	100
	676	232		

表４　JFN における語彙カバー率・意味フレームカバー率

５．考察

調査結果について、意味フレームカバー率、語彙カバー率、JFN 拡張の観点から考察する。まず、意味フレームカバー率は、形状詞が若干低いがおしなべて高いことが判明した。小原(2011)の結果と比較しても上昇している。この結果は基本語の語義については既存の意味フレームである程度カバーできていることを示唆している。[3]

語彙カバー率は、意味フレームカバー率と比べるとかなり低い。読みが同じ漢字のものを採用しているため異表記が理由で語彙カバー率が低かったわけではない。特に、形状詞は、形容詞・動詞に比べて語彙カバー率が格段に低いことが明らかとなった。今後の詳細な分析が必要である。

今回の調査対象は、JFN データベースのうち、BCCWJ 書籍・白書ジャンルのコアデータに対するアノテーション結果であった。今回の調査結果は、基本語彙の定義は意味フレームとしてある程度カバーできているが基本語彙数としては BCCWJ の白書・書籍だけでは足りないことを示唆している。JFN には BCCWJ を対象とした全文テキストアノテーション開始以前に京大コーパス・青空文庫を対象とした語彙アノテーション結果のデータベースも別に存在する。両者のデータベースを統合することにより語彙カバー率を上げ JFN を拡張することを検討すべきである。

６．終わりに

今後の展望としては以下の３つの方向性を検討中である。まず、英語対訳テキスト中の自立語についても FN の語彙カバー率・意味フレームカバー率を明らかにし、JFN のそれらと比較する。次に、今回の調査結果に基づき JFN データベースを拡張する。意味フレームカバー率が比較的高く語彙カバー率が低いということは、新たな意味フレームを定義せずとも、今回未登録であることが判明した語を JFN の該当意味フレームに登録することで語彙カバー率をある程度あげられる、ということである。最後に、JFN アノテーションの難易度の調査である。今回の調査手順３で JFN のアノテーション未経験者が自立語への意味フレーム名付与作業を行い、手順４で JFN アノテーション歴 10 年のアノテータがその検証作業を行った（第３節参照）。手順３における誤りの分析を詳細に行うことにより、JFN アノテーションや意味フレーム定義の難易度を明らかにすることができると考えている。

[3] 今回意味フレームが付与できなかった自立語は 32 異なり語であった。それらのうち日本語教育語彙表（https://jreadability.net/jev/)で初級とされている語はそのうち 12 語であった。

主要参考文献

樺島忠夫(1955).「類別した品詞の比率に見られる規則性」『国語国文』24 巻 6 号, pp.385-387.

樺島忠夫(1963).『表現論――ことばと言語行動』綜芸舎.

山崎誠(2014).「言語単位と文の長さが品詞比率に与える影響」　第五回コーパス日本語学ワークショップ予稿集, pp.233-242.

やなせたかし・たまきゆりこ訳(1991).『英語対訳つき　とべ！アンパンマン１』フレーベル館.

小原京子 (2011). 「BCCWJ への日本語フレームネットの意味アノテーション」,『現代日本語書き言葉均衡コーパス』完成記念講演会，pp. 371-376.