日本語コーパスに対する動詞項構造シソーラスの概念と意味役割のアノテーション

竹内 孔一　　上野 真幸
岡山大学大学院 自然科学研究科
{koichi, ueno}@cl.cs.okayama-u.ac.jp

1　はじめに

文における述語の意味は単語が違っていても共通する意味概念があり，ほとんど同じ意味であることがある．例えば,「社長が太郎を一人前に育てる/仕立てる/鍛えあげる」では，ある対象「太郎」に対して動作主「社長」がある働きかけをして成長させたことを意味している．こうした意味概念をクラスタ (例えば [成長] という意味概念) としてまとめておき，自他の違いなどを整理しておくと，例えば,「太郎が一人前に育つ/成長する」も同じ意味概念として扱うことができる．さらに，前文の「社長が」は「育てる」という動作の動作主体であるのに対して,「太郎を」は「育つ物」であり，それは後文の「太郎が」と同じであるという関係付けができると，文内の要素同士の対応も容易に取ることが出来る．こうした述語間の関係を同定するために，各述語の語義どうしで共通する意味概念でまとめて，その述語に係る要素 (主に項と呼ばれる) に対して意味関係のタイプ分けである意味役割を付与しておく述語項構造辞書が必要である．本研究室では Lexeed 辞書の動詞，形容詞，形容動詞を対象にこれらを人手で分析してまとめ，述語項構造シソーラス (以降シソーラス) として内部でまとめている[1]．シソーラスでは一つの語義に対してほぼ一例文しかないため例文の拡張が，自動付与システムの構築 [5, 4] に必要である．

そこで，本研究では日本語コーパス (以降 BCCWJ)[2] のコアデータに対して，シソーラス (11902 語，概念は 1084 種 (5 階層)) を人手で付与することで各概念の事例の拡張を進めている．例えば,「変わる」の例では

> [同じ車種なら、$_{Premise}$] [普通は $_{領域}$] [ドアの枚数で $_{原因}$][全長は $_{対象}$] 【変わり $_{変更}$】ません。

のようになる．ここで，動詞の意味概念は【】[2]で表し，意味役割は [] で表す．本稿では動詞約 700 種類，約 6500 件について付与する際の作業枠組み，付与結果，ならびにそこから得られた問題点について明らかにする．

2　動詞の意味概念と意味役割の設計

述語項構造シソーラスは動詞項構造シソーラスを拡張して Lexeed[1] に付与されている全ての動詞，形容詞，形容動詞の語義を人手で分析して，例文を作成し，概念をシソーラス形式で整理して，例文に対して，項の分析を行い意味役割を付与した物である．詳細は文献 [6] に譲るとして，意味役割の設計で大きく異なったことだけをここでは押さえておく．(1) 大きな点は従来，起点や着点といった移動に用いられて考えられてきた意味役割は，変化前，変化後の細分類であったとして整理しなおした点．(2) 動作主，原因，手段，理由，causer などは対象となる動詞が自動詞か他動詞か，また項が人か概念か動作か物かの違いであり，外的要因として大きくまとめられる点．この 2 点を整理したことである．

3　動詞概念と意味役割付与

動詞の意味概念と意味役割をどのような手順で付与したか，作業環境と付与結果について明らかにする．

[1]前段階の動詞に対する整理した結果として動詞項構造シソーラス [3] として公開している (http://vsearch.cl.cs.okayama-u.ac.jp/).

[2]シソーラスの意味概念は階層構造 (最大 5 階層) であり，上記の場合は【状態変化あり-変更-変更-変更】の最終分類のみ示す.

3.1 作業内容

シソーラス上に定義されている動詞の事例拡張の観点から作業者の手順を記述する．

ap1 シソーラスで定義されておりかつ BCCWJ にも存在する動詞を選択

ap2 例文を選択

ap3 動詞の語義を選択 (語義無し可)

ap4 係り元の文節や句，文を同定

ap5 係り元動詞の意味役割を選択

上記各項目について，簡単に説明する．まず ap1 では作業システムで動詞のリストと動詞が例文に含まれている数などが付与されている．そのうち，シソーラスに登録されている動詞を選択する．

次に ap2 ではシソーラスでの意味概念が 3 以下なら 10 例文，4 種以上なら 20 例文として例文を選択する．例えば「測定する」などは 1 概念しかなく,「上がる」ならば 15 概念ある．例文選択は，本来ならば必要とする概念を幅広く獲得すべきであるが，事例付与の最初の段階であり，時間をかけ 1 ると付与がほとんど出来ない恐れがある．簡単な事例でもまず量を出す必要から，こうした簡単化した指示を行った．現段階では人手で行ったが，統計的手法などを利用した補助システムが必要であろう．

ap3 では BCCWJ の長単位 (複合動詞や「〜する」などが 1 単位) に対して動詞の語義を選択して付与する．この際，シソーラスの事例に無いと作業者が判断した場合は' 概念無し' を選択する．例えば「抱える」の場合，シソーラスの事例として「腕に抱える」【掴む】,「パートを抱える」【着任】,「借金を抱える」【請負】,「発電所を抱える」【維持】の 4 つの概念が定義されているが,「頭を抱える」の概念は「抱える」という動詞には登録がない．この場合，作業者は語義無しを選択する．しかしながら，概念体系 1048 の中には当然これに対応する概念が存在する．この場合心理的変化を表す「困る」と同様の意味であり【苦しみ】という概念を全体から見つけて付与すべきである．しかし，作業者にはコストが高い作業であり，量を出す方針から，これらの作業は管理者が行うという立場で，この作業は省略している．

ap4 では動詞に係る要素に対する意味的な関係 (意味役割) を付与するために，係り元を同定する．BCCWJ には係り受け解析結果などが付与されておらず，また複文や並列句など構造が簡単ではなかったため人手で付与した．

> 一度ＤＴＰソフトなどで編集しておけば、[その都度 $_{Time-Repeat}$] [新しい作品を $_{対象}$]追加する などして印刷して総じて，常に最新版を作れます．

上記の文では複文の従属節内に付与する「追加する」の係り元を同定している例である．文の途中であるため,「〜しておけば,」という条件節は主節の「作れます」に係るため係り元とは判定しない．

最後に ap5 では 95 種類の中から意味役割を付与する．作業者に対しては 95 種類のタグが全て提示されてその中から選択するため，容易ではないが，シソーラス内に定義されている各動詞の概念の意味役割の付与事例を提示するので結果を見る限り混乱はしていない．

管理者は上記の作業を確認しつつ，語義の不足や問題点などを記録しながらマニュアルの更改や意味役割の整理を行っている．

3.2 作業環境

作業期間は 2012 年 9 月から 2013 年 1 月末までであった[3]．まず作業をする場所として管理者と作業者を同じ部屋で作業することにした．またアノテーション作業システムを構築してブラウザベースで作業できるようにした [7]．作業者は 3 人の学部学生 (所属は文系 2 名理系 1 名) と 1 人の大学院生であり，雇用期間もバラバラである．大学院生を除いて全員意味分類の経験は無く，語義と意味役割について各個人 3 時間程度の説明を行い付与させた．

作業者は学生であるため講義のある時間は作業が出来ない．よって作業時間がまちまちでアノテーションの全体会議を行うことは不可能であった．よって作業システムとしては家からでもできるが，管理者と同じ部屋で開発することで質問などすぐに対応する形式をとった．

さらに管理者は付与結果を作業システムを通して確認し，付与がずれている作業者に対しては説明を加えて理解を安定させた．作業状況からする特徴は (1) 具体的な分類に関する質問はあまりしない，(2) アノテーションで揺れたところを自分で管理して報告するといったことはできない，(3) 作業システムに関する不満も「こういうものだ」と作業者が思い込んでしまい (管理者が先生であることも理由かも入れない) 改

[3] よって本稿ではまだ作業中であるため不確定の部分がありつつ原稿を記述している．

表 1: 各作業者の概念タグの付与量と一致率

	A	B	C	D
(1)	2264	6118	2037	6339
(2)	124	248	117	263
	(5.4%)	(3.8%)	(5.7%)	(4.1%)
(3)	87.3% (6584/7546)			

善点の提案は無かった．よって管理者は作業者の行動とアノテーション結果から何が必要かを常に分析する必要があった．

しかし，こうした作業環境の構築で下記に示すように速いスピードでのタグ付与が実現できたと考えられる．

3.3 付与結果

表 1 に動詞の意味概念の付与数と一致率について示す．作業者は 4 人 (A, B, C, D) おり，１つの事例に対して 2 人以上の作業者が付与することにしている．そこで一致率はある事例に対して複数作業者がタグを付与した場合に，誰かが一致すれば一致と見なして計算する．ここで (1) 付与数，(2) 該当する’概念無し’の付与数，(3) 概念が 2 人以上で付与されている場合に，2 人の作業者の概念タグが一致した箇所，を示している．

まず (2) の付与結果からシソーラスの概念分類の BCCWJ に対する非カバー率が多くて 6%以下と考えられる．シソーラスに掲載されていない動詞は対象外にしているので，掲載されていれば語義は 94%程度の割合で定義されていることを示している．当然この結果は，上述したように，1084 分類の全てを見ているわけでないのでカバー率はこの値より高くなることが予測される．

また概念タグの一致率は 87.3%であった．まだ作業中であるので最終の値ではないがシソーラスの定義と事例による付与は専門家でなくてもある程度一致すると考えられる．

一方，意味役割の方は構築しながら体系の整理を行っているため，揺れている部分があるが，それでも表 2 の結果を得た．(1) は各作業者の意味役割付与箇所の数を示しており，(2) は意味役割の付与部分が 2 人以上一致している場所で，2 人以上の作業者が同じ意味役割タグを付与した一致率を表している．95 種類あるなかから 8 割を超えるタグの精度が一致していることは作業が成功していると見える．ただし，管理者によるチェックが進んでいないため最終的に一致した程度が管理者との判断でどの程度ずれているかはまだ明確ではない．

表 2: 各作業者の意味役割タグの付与量と一致率

	A	B	C	D
(1)	316	10753	2597	12241
(2)	81.8% (8199/ 10024)			

4 アノテーションの考察

アノテーション作業で語義概念と意味役割の付与において問題になった点を挙げる．

使役形や受動形，テイル形などで現在形と意味が透過的でない

今回の付与の枠組みでは動詞の概念を現在形で設定し，使役形や受身形は基本的に変わらない範囲で記述する．しかし今回の付与で，受身形や使役形で別の概念と考えられる事例がいくつか見つかり，今のところ’概念無し’を付与している．例えば「知る」の場合「言葉の意味を知る」【状態変化-理解】,「合気道を知る」【状態変化-習得】,「議院を知る」【状態-結束】などであるが，下記の例はこれらとは異なる意味である．

> これは、脱構築という言葉で知られているフランスの思想家ジャック・デリダの文章です。

この場合「噂が広まる」などと同様の【普及】という概念が当てはまる．これは基本形の「知る」では出てこない概念であり，活用形まで含めて語の意味の異なりを扱う枠組みが (全てではなくても) 必要であることがわかる．同様に使役形の場合も「幅を利かせる」のように,「幅が利く」と基本形では意味を成さず必ず使役形でのみ現れる概念が存在する．どの程度このような語があるかは明らかでないが MWE と考えられる．

複合動詞の一部の場合の扱い

構成的な複合語の場合は概念が付与できる場合もあるが，非構成的な複合語の場合には’概念無し’とした．例えば，構成的な複合語の場合,「水筒を持って行く」の「持つ」であれば,「手に持つ」という意味で【携帯】という概念があり，複合動詞内の要素でも「持つ」が取る概念を付与することができる．しかしながら下記の「持って行く」では【携帯】の意味は感じられない．

アメリカ映画や、〜中略〜は創作ですので感情の高ぶりはハイレベルまで 持っていく よう描かれていると思います。

この場合は「持つ」は独立した意味は無く,「持っていく」で一つの意味概念【程度の変化】とするのが妥当に思える．よって現段階ではこの例のような「持つ」には’概念無し’としている．

作業者が理解していない

作業者が概念の階層性など理解できずにあやまる例が見受けられた．典型的と思われる例を下記に挙げる．

- 提示されている概念を理解せずに付与する
  例えば「壊す」の概念には物理的な物を壊す【破壊】と概念的な計画などを無くす【消滅】の意味を儲けているが「試合を壊す」を物理的な【破壊】にする作業者が多かった．しかし，壊す対象が抽象的な概念であり，しかも「勝つ機会の消滅」という意味と考えられるので【消滅】と付与するのが正解である．

- 例文の類似性に引っかかる
  例えば「受け取る」の動詞に対して付与対象の例文が「報告を受け取る」の場合，概念としては【入手】が正しい．しかしながら,「受け取る」には「忠告を B と受け取る」【判断】の概念があり,「報告」と「忠告」の文字的な近さから【判断】を付与するという物である．これはシソーラスの概念体系が変化において，物理的な移動と人間の理解に関する変化をするどく分けているという背景が伝わっていないため起こっていると考えられる．この理解は容易ではないので，作業として正確さを求めるには例文の工夫が重要であることが伺える．

- 前に付与した判定を直さない (作業の一貫性の欠如)
  例えば前の方で付けた例文の判定と後の方で付与した例文の判定が変わってしまっているのに，そのままタグを付与して次の作業を行う場合がある．これは作業者自身へのインタビューでは本人は一貫性を保とうと努力しているにもかかわらず起こる例で，おそらく概念付与の作業自体が難易度が高いと考えられる．

5 おわりに

日本語コーパスに現れる動詞について述語項構造シソーラスの意味概念と意味役割の付与を行った．約 5ヶ月の作業で 2 人作業者の付与結果が一致した件数は，動詞の概念では 6584 件，意味役割では 8199 件の件であった．現段階では概念タグのゆれや枠組みについて整理しつつある段階で，今後意味役割についての分析をすすめて，意味役割の関係タグを精緻化する予定である．

謝辞 本タグ付きコーパスの構築に当たって国立国語研究所にご支援いただいた．ここに記して感謝する．

参考文献

[1] Sanae Fujita, Takaaki Tanaka, Fransis Bond, and Hiromi Nakaiwa. An implemented description of japanese: The lexeed dictionary and the hinoki treebank. In *COLING/ACL06 Interactive Presentation Sessions*, pp. 65–68, 2006.

[2] Kikuo Maekawa. Balanced corpus of contemporary written japanese. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pp. 101–102, 2008.

[3] Koichi Takeuchi, Kentaro Inui, Nao Takeuchi, and Atsushi Fujita. A thesaurus of predicate-argument structure for japanese verbs to deal with granularity of verb meanings. In *The 8th Workshop on Asian Language Resources*, pp. 1–8, 2010.

[4] Koichi Takeuchi, Suguru Tsuchiyama, Masato Moriya, Yuuki Moriyasu, and Koichi Satoh. Verb Sense Disambiguation Based on Thesaurus of Predicate-Argument Structure. In *Proc. of the International Conference on Knowledge Engineering and Ontology Development*, 2011. 208–213.

[5] 竹内孔一, 土山傑, 守屋将人, 森安祐樹. 類似した動作や状況を検索するための意味役割及び動詞語義付与システムの構築. 電子情報通信学会，言語理解とコミュニケーション研究会, NLC-2009-33, pp. 1–6, 2009.

[6] 竹内孔一, 竹内奈央, 石原靖弘. 述語の分析に基づく文書解析の考察. 情報処理学会自然言語処理研究会, NL-207-15, 2012.

[7] 上野真幸, 竹内孔一. 動詞語義及び意味役割付与作業システムの構築. 第 2 回日本語コーパスワークショップ, NL-207-15, pp. 69–76, 2012.