社会的状況下の言語使用を踏まえた日本語学習支援への取り組み

理学専攻・情報科学コース　劉 ボセン

1　はじめに

日本に住む外国人の日本語教育に対するニーズは年々高まっているが，さまざまな事情で学習の機会を得られない人が少なくない．このような背景から，自然言語処理技術を用いて外国人学習者の独学環境の整備や，過疎地の日本語教師への支援が必要だと考えられる．

日本語における言語使用は，他の言語よりも話者間の相対的地位や社会的状況を強く反映しているとされる [1]．そのため，様々な自然言語処理タスクにおいて日本語テキストを処理する際，より正確なモデルを構築するためには，テキストを取り巻く社会的状況を考慮することが必要になる．モデルが個々の社会的状況を適切に捉え，学習するためには，それら社会的状況に関する属性情報を含むコーパスの存在が，より実用的で効果的なモデルの開発に役立つと考えられる．本研究では，社会集団における言語使用の観点から言語分析を行う選択体系機能言語学に基づき，より詳細な社会的要因に関する分析情報を含む日本語コーパスの作成と検証を行ったうえ，日本語学習支援システムの構築を試みる．

2　社会的状況の定義

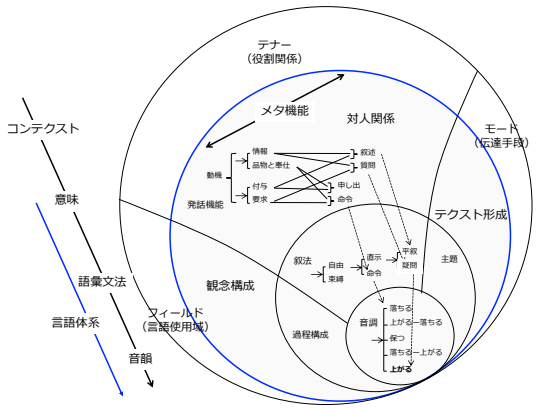

図 1: SFL による言語体系（図は [2] から引用）

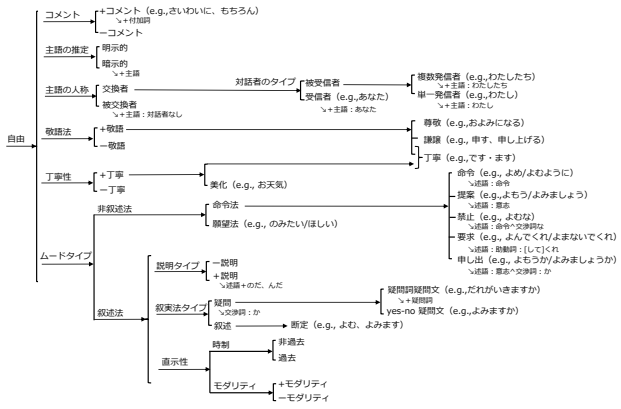

図 2: 叙述の選択体系網（図は [3] から引用）

選択体系機能言語学（Systemic Functional Linguistics, SFL）では，言語体系は，意味層，語彙・文法層，表現層というそれぞれ異なる種類の記号体系が同心円的に階層性を成し，コンテクストによって包括されているとされ，社会集団の価値や共通の認識に基づいた状況下における言語使用を表現する包括的なモデルとなっている [4]．SFL にによって示される言語体系を図 1 に示す．コンテクスト層は，言語の使用域を示す「活動領域（フィールド）」，話者間の社会的関係を示す「役割関係（テナー）」，使用する媒体を示す「伝達様式（モード）」といった 3 つの特性の下，状況を定義している．言語体系には，それらコンテクストの 3 つの特性それぞれに対応する観念構成的意味，対人関係的意味，テキスト形成的意味の 3 つのメタ機能が働いており，選択体系網からの言語資源の選択に制約を与えることによって，状況に適した発話が形成される．本研究では，言語生成時に含まれる隠れた情報を得るために，各文に対して，上記 3 点のコンテクスト要素を含む注釈を付与している．とくに，取り扱う敬語は，対人関係的意味を反映して語彙・文法層内に記される叙述の選択体系網から選択された素性に基づき表出される．図 2 に叙述の選択体系網を示す．

3　SFL を用いたコーパスの構築

本研究では，上述した SFL によって捉えられる社会的状況を考慮したコーパスを構築する．

3.1　KeiCO コーパス

日本語敬語コーパス「KeiCO コーパス」の構築において，敬語に関する書籍 [5] やインターネット上の記事などから敬語表現を含む原文として収集し，クラウドソーシングにより日本語母語話者のアノテータ 40 名に，SFL の素性を注釈として敬語のレベルの付与を依頼した．各アノテータには約 75 の原文が割り当てられ，原文の意味を維持しつつ，できる限り他の敬語レベルに書き替えてもらった．書き換えが難しい箇所は空欄のままを許可する指示をした．また，書き換え後の文に指定した敬語のレベルを参考にレベルスコアを付けてもらった．このように作成したコーパスは 9,524 文からなり，敬語を処理する機械学習や統計解析のためのコーパスとして，また，日本語学習用教材などとして利用可能である．

表 1 に，KeiCO コーパスの概要を示す.KeiCO コーパスでは，各文に対して，敬語のレベル，書き言葉，話し言葉，尊敬語，謙譲語，丁寧語と言語使用域（フィールド）の 7 種類の注釈が付与されている．1 行目は SFL の叙述に関する選択体系網の属性が注釈として示されている．1 列目のコーパス文に対して，それぞれの属性は，0 または 1 の値が付与されている．1 はその属性に該当することを示し，0 はその逆を示している．

3.1.1　実験と考察

KeiCO コーパスを用いて，コーパス内の注釈に対する分類精度を検証した．分類モデルは，東北大学で構築された汎用日本言語モデル BERT_{BASE}[1]を使用し，

[1]https://huggingface.co/cl-tohoku/bert-base-japanese

表 1: KeiCO コーパスの例

KeiCO コーパス文	敬語レベル	書き言葉	話し言葉	尊敬語	謙譲語	丁寧語	活動領域
本日は, かねてより相談したいことがあり, 参上しました.	1	1	0	0	1	0	相談

表 2: KeiCO コーパスにおける各素性の分類精度 (10 回平均)

分類精度	敬語レベル	書き言葉	話し言葉	尊敬語	謙譲語	丁寧語
データ量 total	0.73	0.59	1.00	0.81	0.91	0.84

表 3: コーパスの概要：従業員が顧客に対してお知らせする場面の例

場面	本文	対話参加者							発話機能			
		上下関係(受信者)	上下関係(送信者)	送信者身分	受信者身分	内外関係	送信者数	受信者数	送信者の動き	送信者の動き(詳細)	やりとりにおける役割	やりとりされるもの
顧客 A 様からご依頼頂きました商品が入荷しました。あなたは A 様に商品の入荷をお知らせするメールを書いて下さい。	件名: 商品入荷のお知らせ A 様 日頃より弊社の製品をご愛顧くださり、まことにありがとうございます。先月ご注文いただきました商品ですが、本日入荷いたしましたので、ご都合のよろしいときにご来店いただければと存じ上げます。 ご自宅へのご配送をご希望される場合は、お手数でございますが、下記の配送センターまでご連絡くださいますようお願い申し上げます。 配送センター 担当 XX	目上	目下	従業員	顧客	外	個人	個人	質問	問い合わせ	与える	情報

KeiCO コーパスを用いてファインチューニングされ作成されている．KeiCO コーパス全体を学習データ，検証データ，評価データに 6 : 2 : 2 の割合で分割し，エポック数を 30 とした．表 2 に，KeiCO コーパスの各素性に対する分類精度を表す. 結果として，話し言葉，尊敬語，謙譲語，丁寧語は高い分類精度を収める一方で，敬語レベル，書き言葉はやや低い精度になった．敬語のレベルについては細かく 4 つに分類されるため，他の 2 値分類タスクより精度が劣ったと考えることも自然である. また，書き言葉と話し言葉の境界は曖昧であり明確な注釈づけは難しい. 分類精度向上のためには，複数のアノテータのコンセンサスを用いるなど工夫が必要だと考える．

3.1.2 敬語自動変換システムの構築

KeiCO コーパスでは，尊敬の程度をテナーで反映した 4 つのレベルを設定しているため，モデルに敬語の度合いを学習させることが可能である．本研究は日本語 T5 事前学習済みモデル[2]を利用し，学習データを全く敬語を使わない文（敬語レベル 1）に，教師データを同じ意味の敬語文（敬語レベル 4）に設定し，日本語学習者の敬語の使用を支援する敬語自動変換システム（https://deepjapanese.com）をオンラインから利用できるように構築した．

3.2 日本語ビジネスメールコーパス

KeiCO コーパスは 1 文しか分析しておらず，SFL の階層的構造を示していないため，その拡張版として，送信者と受信者の社会的立場の違いに基づいた書面でのコミュニケーションという会話的な要素を持つ電子メールというジャンルを選択し，社会的状況に関する情報が付与された日本語コーパスの構築を試みた．表 3 は，コーパスの全体像の例を示されている．アノテーションに関して，テナー（役割関係）の選択体系網を元に，「対話参加者」及び「発話機能」で決められている．このコーパスは，770 種類の社会的状況場面を含んでおり，各場面毎に 5 通のメールを提供し，合計 3,850 通のメール情報がある．

4 おわりに

本研究は，選択体系機能言語学に基づき，日本語テキスト内の社会的地位の情報を反映した日本語コーパスを作成した．作成されたコーパスは，アノテーションのタグとして用いられる SFL の選択体系網の選択肢をすべて使っているわけではなく，社会的役割関係を重視したものとなっている．今後の課題として，作成したコーパスの機械学習課題における利用と性能評価を行うとともに，SFL での状況のコンテクストからテキストが具現される過程を捉えたコーパスアノテーション手法の確立を目指すつもりである．

参考文献

[1] 松村瑞子・因京子. 日本語談話におけるスタイル交替の実態とその効果. 『言語科学』, No. 33, pp. 109–118, 1998.

[2] 小林一郎. 意味へのアプローチ：ハリデー言語学の観点から, 2017.

[3] 角岡健一［編著］, 飯村龍一, 五十嵐海理, 福田一雄, 加藤澄. 機能文法による日本語モダリティ研究. くろしお出版, 2016.

[4] M. A. K. Halliday and C. M. I. M. Matthiessen. *Construing Experience Through Meaning: A Language-Based Approach to Cognition*. Continuum, 1999.

[5] 坂本達, 西方草志. 敬語のお辞典. 三省堂, 2009.

[2]https://huggingface.co/sonoisa/t5-base-japanese