選択体系機能言語学に基づく日本語敬語コーパスの構築と検証

Muxuan Liu　　小林一郎
お茶の水女子大学
{liu.muxuan,koba}@is.ocha.ac.jp

概要

日本語は，話者間の社会的地位によって発話文の表現が異なり，それは敬語の使い方の違いに顕著に現れている．日本語は，他の外国語と違い多くの敬語の種類が存在し，機械翻訳や対話システムにおいて，その意味の違いを正しく処理を行うことはとても重要である．しかし，社会的立場を踏まえて敬語を取り扱うようなコーパスは未だ存在しない．本研究では，このような背景を踏まえ，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選択体系機能言語学に則り，社会的地位関係の情報を含む敬語コーパス（KeiCO コーパス）を構築した．また，BERT を用いた識別課題を行い，構築したコーパスの性能を検証した．

1　はじめに

現代日本語には，「-なさる」「お/ご-なさる」などの尊敬語や，「お/ご-する」「お/ご-申し上げる」などの謙譲語が存在する．このような敬語の表現に着目した機械処理の研究はすでに多く行われている [1–5]．敬語には決まった文法変換規則があるため，この文法規則を用いて簡単な敬語の訂正や補助的なシステムを作ることが可能である．しかし，現実生活においての敬語の使い方は複雑であり，敬語の生成は，話者間の社会的地位，親密さ，場面などを考慮する必要がある．敬語の使用について，大規模コーパスを用いて多くの用例から敬語使用の実態を分析するのが一般的だが，実際，言語使用域，対話者間の社会的な役割関係，対話の手段など，社会集団の言語使用に関する詳細な情報を含むコーパスが存在しない．それゆえ，社会的要因を考慮して適切な敬語を使用する機械学習モデルを構築することは容易ではない．

そこで本研究では，これらの問題点を踏まえ，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敬語コーパス（**KeiCO コーパス**）の構築と検証を試みる．また，構築した KeiCO コーパスを一つの言語資源として公開する予定である．

2　選択体系機能言語学

選択体系機能言語学（Systemic Functional Linguistics, SFL：詳細は付録 A.1 を参照）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（付録：図 2 参照）．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 1 に叙述の選択体系網を示す．本研究では，この選択体系網に定義される素性を，敬語コーパスへの注釈として用いる．

3　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 [6] 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

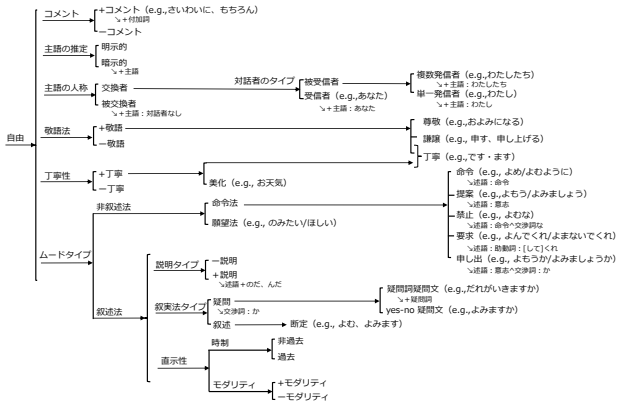

図 1: 叙述の選択体系網

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．上記の作業を終えた後，さらに 20 名の日本語母語話者のアノテータに注釈の正しさを検証を依頼し，明らかな誤りは手作業での修正を求めた．

このように作成したコーパスは 10,007 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．1 行目は SFL の叙述に関する選択体系網の素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，それぞれの注釈は，0 または 1 の値が付与されている．1 はその注釈に該当することを示し，0 はその逆を示している．各注釈の詳細な定義を，3.1 節に示す．

3.1 構成とアノテーション

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．詳細な定義を以下に示す．

3.1.1 敬語のレベル

敬語の選択は，主として役割関係（テナー）に反映して行われる．テナーには，社会的地位による上下関係（例，上司と部下，先生と生徒，等）や人間関係の親密さ（例，友人，知り合い，等）といった社会的対人関係も含まれる．

これにより，コーパス中に表現される相手への尊敬の程度は，テナーによって規定可能であると考える．KeiCO コーパスでは，尊敬の程度をテナーを反映した 4 つのレベルを設定した．各レベルを以下のように定義する．

Level 1：最高レベルの尊敬度　ニュースや，非常にフォーマルな講演，正式的なビジネスメールなどでよく使われる敬語のレベルである．最高レベルの尊敬度に属する文では，日本語の文法規則に従い，動詞が尊敬や謙譲を表す形に変形されることや，それ自体が尊敬の意味を持つ言葉が使われることが一般的である．また，尊敬語と謙譲語を組み合わせた敬語連結の形にもなり得る．

Level 2：第二レベルの尊敬度　ビジネスレター，一般的な学術・ビジネス講演，サービス業などで広く使われる．文法規則に従い，動詞が尊敬や謙譲を表す形に変形されるが，敬語連結の形が少ない．

Level 3：第三レベルの尊敬度　複雑な動詞変形は使われず，殆どが丁寧語や美化語のみ使われる．

Level 4：第四レベルの尊敬度　敬語は全く使用されない．レベル 3 よりもカジュアルな表現で，美化語や省略語，ネット用語が登場することもある．

3.1.2 書き言葉・話し言葉

敬語はテナーの影響のみならず，伝達様式（モード）の影響により，選択される表現が異なる．モードには，広義的に多様な伝達媒体によるものもその範疇に含まれ，SNS，電話，メールなどが挙げられるが，狭義的に話し言葉と書き言葉だその範疇とみなされる．KeiCO では文章の敬語校正などのタスクにも適応性があるように，モードに対する素性を書き言葉と話し言葉と定義し注釈付を行なった．一般に，書き言葉は「だ・である調」，話し言葉は「です・ます調」であると言われている．しかし，現実の場面，例えばビジネスメールの作成などでは，そのような決まりはない．様々な用途へのコーパスの適応性を高めるために，KeiCO コーパスでは書き言葉に略語や方言を含まず，文法規則に従い，文語的な表現が多い言葉と定義した．また，話し言葉は略語や方言，感嘆詞なども含んだ口語的表現が使われる言葉と定義した[1)]．

1)　森山 [7] によれば，書き言葉と話し言葉の境界は曖昧であり，人それぞれ定義が異なるという．実際に，日本語母語者にとっても，外国人日本語学習者にとっても，適切な使い分けは容易なことではない．

表 1: KeiCO コーパスの概要

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
誠に遺憾でございます.	1	1	1	1	0	0	mail
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
今日は，折り入ってご相談したいことがあって伺ったのですが.	2	1	1	0	1	0	相談
今日は相談したいことがあったため，来ました.	3	1	1	0	0	1	相談
今日はずっと相談したいことがあって来た.	4	0	1	0	0	0	相談
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3.1.3 尊敬語・謙譲語・丁寧語

SFL の叙述に関する選択体系網の素性に基づき，尊敬語，謙譲語と丁寧語の 3 つの尊敬表現を注釈として利用する．尊敬語は，話し手が話題の主体である人物を尊重する意思を表現し・尊敬される人の行動，物，名前にも使われる．謙譲語は，話し手が自分の言動を一段下に下げ，聞き手に敬意を示す意思を示す．丁寧語は，主に「です・ます」など助動詞の語尾編かを促し，話題を美化し，言葉に敬意を込めるために使われる．

3.1.4 活動領域

活動領域（フィールド）は，言語の使用域を示し，会話の場面，あるいは，話題を意味する．敬語は，ビジネス文書や講演会など特定の活動領域においてその使用が影響される．そのことを考慮し，KeiCO コーパスでは注釈に具体的な活動領域を示す注釈を付与している．現在，KeiCO コーパスでは 122 の種類の活動領域をその選択肢としている．(付録：表 5 参照)

4 KeiCO コーパス解析

KeiCO の特徴的統計量を表 2 に示す．

4.1 統計的分析

KeiCO おける語彙の使用について，最も基礎的な情報である延べ語数と異なる語数を確認し，K 特性値を求めた．

K 特性値 (characteristic K) は，ユール [8] によって提案された語彙の豊富さを示す指標であり，数値が小さいほど語彙の豊富さが高いことを示す．K 特性値は, 単語の出現頻度がポアソン分布に従うと仮定している. いま, 延べ語数が N, 異なり語数が V である文章の中に，m 回出現した単語数を $V(m,N)$ とした時，K 特性値は式（1）で定義される．

$$K = 10^4 \times \frac{\sum_{\text{all } m} \left[m^2 V(m,N)\right] - N}{N^2} \tag{1}$$

KeiCO において，敬語のレベルの上昇とともに K 特性値は高くなるが，レベル 2 の短文数が他のレベルに比べ少ないため，レベル 2 は他のレベルより語彙の豊富さが高いと言う結果となった．また，語彙に関する使用について，硬い印象のある漢語に注目し，1 文に含まれる漢語の平均数を求めた．結果として，敬語レベルの上昇により漢語の使用量が増えていくことが確認できた．

4.2 敬語素性の分類精度

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}[2] を使用し，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 3 に，KeoCO コーパスの各素性に対する分類精度を表す．コーパスから $0.01, 0.1, 1(total)$ の割合（約 100 文,1000 文,10000 文）でランダム抽出し，データ量から分類精度への影響を確認する．

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．また，データ量が 10 倍に増加する平均精度増加率について，敬語レベル，尊敬語，丁寧語は高い一方，書き言葉，話し言

2） https://huggingface.co/cl-tohoku/bert-base-japanese

表 2: レベル別の統計結果

敬語レベル	短文数	平均文長	単文平均漢語数	延べ語数	異なり語数	k 特性値
Level 1	2584	18.2	2.6	47111	4744	135.70
Level 2	2046	16.4	2.1	33476	3897	136.23
Level 3	2694	15.2	1.8	40980	4448	130.28
Level 4	2683	13.5	1.6	36233	4315	129.80
Total	10007	15.8	2.0	157806	6465	125.54

表 3: 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 0.01	0.482	0.646	0.990	0.600	0.894	0.706
データ量 0.1	0.653	0.686	0.952	0.780	0.887	0.810
データ量 total	0.727	0.698	0.948	0.816	0.906	0.842
平均精度増加率	23.4%	3.9%	-2.1%	17.3%	0.7%	9.4%

葉，謙譲語は非常に低い増加率になった．その中，特に話し言葉は負の平均精度増加率になっている．

尊敬語，丁寧語　尊敬語，丁寧語について，分類精度，平均精度増加率が高い理由は，文法的特徴が文中に表出しているため，特徴の識別が容易であったためと考える．

話し言葉，謙譲語　今回，コーパス中，話し言葉と謙譲語が一方のラベルに偏っているため (付録：表 4 参照)，分類モデルがうまく学習できなかったことが平均精度増加率が低い原因に繋がったと考えている．

敬語レベル　3.1.1 節で述べたように，敬語のレベルは細かく 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．また高い精度増加率はコーパス中のレベルがバランスよく数が揃えられ，タスクの精度向上に貢献できたとも言える．

書き言葉　3.1.2 節で述べたように，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．

5　おわりに

本研究は，選択体系機能言語学に基づき，話し手と聞き手の社会的地位の情報を反映した，日本語敬語コーパス「Keico コーパス」を作成した．KeiCO コーパスは，様々な活動領域の下での話者間の社会的役割を踏まえ，伝達様式も考慮して，コーパスに注釈付がされており，汎用性の高い言語資源として，機械翻訳の精度の向上，日本語作文の自動評価・自動修正や文体変換など，様々なタスクに役に立つことを期待している．今回，未対応の課題として，(1) 各ラベルの短文数が均衡ではない点，(2) 書き換えた文の中には，敬語のレベルに合わせた難易度となる適切な語彙に変換されなかったものもあるため，語彙の豊富さに関しては再検討の余地がある点，などが挙げられる．今後，KeiCO コーパスの短文数を増やし，名詞の書き換えなどを重点的に行うつもりである．

参考文献

[1] 李国慶, 吉野孝. 外国人向け敬語文理解支援システムの開発. 電子情報通信学会技術報告, pp. 7–12, 2015.

[2] 飛鳥井元晴, 岸義樹. 敬語文章変換システムの作成. 第 77 回全国大会講演論文集, Vol. 2015, No. 1, pp. 177–178, 2015.

[3] 徳丸瑞稀, 川村華峰, 岡村奈々花, 仲山友海, 中野美由紀. ルールベースに基づくビジネスシーンにおける敬語変換手法の検討. 情報処理学会第 82 回全国大会, pp. 409–410, 2020.

[4] P Resmi and C Naseer. A deep learning approach for polite dialogue response generation. In **proceedings of the International Conference on Systems, Energy Environment (ICSEE) 2019**, August 16, 2019.

[5] Tong Niu and Mohit Bansal. Polite dialogue generation without parallel data. **CoRR**, Vol. abs/1805.03162, , 2018.

[6] 坂本達, 西方草志. 敬語のお辞典. 三省堂, 2009.

[7] 森山卓郎. 話し言葉と書き言葉を考えるための 文法研究用語・12 (特集 スキル話しことばと書きことば–新・言文一致のエクササイズ). 国文学 解釈と教材の研究, Vol. 48, No. 12, pp. 15–22, oct 2003.

[8] George Udny Yule. **The Statistical Study of Literary Vocabulary**. Cambridge: At the University Press, 1944.

[9] 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. 認知科学, Vol. 24, No. 1, pp. 8–15, 2017.

[10] 角岡賢一, 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 (龍谷大学国際社会文化研究所叢書). くろしお出版, 2016.

A　付録

A.1　選択体系機能言語学

選択体系機能言語学（Systemic Functional Linguistics, SFL）は，文化人類学者の Malinowski の考えにロンドン言語学派の Firth が影響を受け，Firth に師事した M.A.K.Halliday によって確立された言語理論である．他の言語学と SFL との大きな違いは，多くの言語学が多様な意味を包括的に扱うことを避け，言語の意味の取り扱いを限定し，文法の側面に焦点をあてるのに対して，SFL はその理論の中に社会集団の中における文化的背景までを含むコンテクストを導入し，社会の中における言語の機能面から言語体系の考察を行なっていることである．SFL にによって示される言語体系を図 2 に示す．言語体系の各層は選択体系網と呼ばれる選択肢からなるネットワークによって言語資源に関する選択の制約が表現されている．また，階層間は実現規則（realization statements）と呼ばれる制約条件によって有機的に連結されている．SFL による選択体系網を使った言語資源の体系化とその選択の手続きが，そのまま文生成のアルゴリズムとして適用可能とみなされ，1980 年代には「システミック文法」と呼ばれて自然言語文生成の主要な言語理論として用いられた．

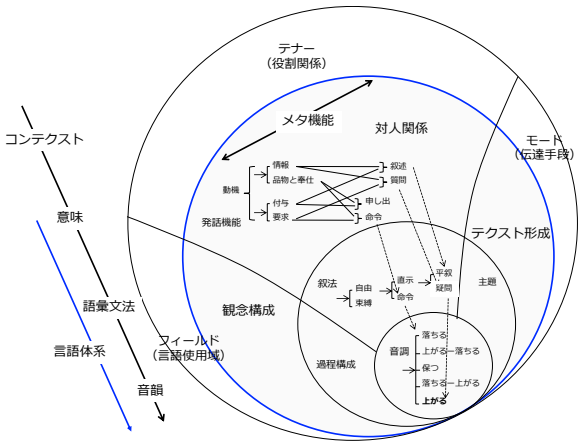

図 2: 選択体系機能言語学による言語体系

表 4: KeiCO コーパスにおける各素性の割合

敬語レベル 1	敬語レベル 2	敬語レベル 3	敬語レベル 4	書き言葉	話し言葉	尊敬語	謙譲語	丁寧語
26%	20%	27%	27%	63%	95%	39%	9%	24%

表 5: KeiCO コーパス内の活動領域一覧

順位	活動領域	数	順位	活動領域	数	順位	活動領域	数	順位	活動領域	数	順位	活動領域	数
1	メール	527	26	楽しむ	112	51	計算	60	76	帰る	56	101	自覚	49
2	食	329	27	支配	105	52	季節	60	77	探す	56	102	作品	49
3	金	326	28	好き	103	53	忠告	60	78	店	56	103	願う	49
4	客	234	29	書く	101	54	応募	60	79	勧める	56	104	裁く	49
5	買う	229	30	仕事	100	55	聞く	59	80	わかる	56	105	信じる	49
6	態度	227	31	祝う	80	56	集まる	59	81	頼む	56	106	叱る	48
7	謝る	217	32	怒る	74	57	見る	59	82	いる	56	107	服を仕立てる	47
8	会う	173	33	手紙	66	58	アイデア	59	83	お参り	55	108	調べる	46
9	贈答	162	34	言う	64	59	検討する	59	84	伝える	55	109	褒める	46
10	挨拶	159	35	赤ちゃん	62	60	体格	59	85	体	55	110	安心	46
11	質問	158	36	遊ぶ	62	61	研究	59	86	別れる	55	111	登場	45
12	政治的講演	158	37	招待	61	62	スポーツ	58	87	断る	55	112	歩く	43
13	言葉	156	38	暮らし向き	60	63	獲得	58	88	体験	55	113	行く	43
14	家	136	39	驚く	60	64	嫌う	58	89	ひま	54	114	ねぎらう	40
15	お知らせ	134	40	勝つ	60	65	反論	58	90	助ける	54	115	確認	40
16	受付	120	41	計画	60	66	逃げる	58	91	お礼	54	116	励む	40
17	関係	120	42	遠慮	60	67	管理	58	92	心	53	117	急ぐ	37
18	作業	120	43	送る	60	68	する	58	93	準備	53	118	恥ずかしい	37
19	公	119	44	契約	60	69	付き合う	58	94	返す	53	119	反対	36
20	関心	117	45	着る	60	70	待つ	57	95	報告	52	120	話す	36
21	秘密	116	46	体調	60	71	取り込む	57	96	病状	52	121	改める	31
22	席	116	47	選ぶ	60	72	任せる	57	97	心配	51	122	紹介	24
23	電話	116	48	教える	60	73	困る	57	98	謙る	51			
24	学校	116	49	あいづち	60	74	終わり	57	99	知っている	50			
25	死ぬ	114	50	相談	60	75	がっかり	57	100	訪問	50			