文のどこにキャラクタ性を埋め込む自由度があるか

刀山 将大† 佐藤 理史† 松崎 拓也† 宮崎 千明‡ 平野 徹‡ 松尾 義博‡
†名古屋大学大学院工学研究科
‡日本電信電話株式会社 NTT メディアインテリジェンス研究所

1 はじめに

すべての文の背後には書き手がおり，すべての発話文の背後には発話者がいる．これらの作成者の痕跡は，なんらかの形で文に埋め込まれていると考えるのが自然である．たとえば，それぞれの小説家に固有な文体は，そのひとつの典型例である．

発話文ではその傾向が顕著であり，発話者がどんな人物であるかを容易に推測できる場合が多い．たとえば，以下の例を考えよう [1].

(1) 「ぼくがだれだか，ご存じなんですか」
「よしてよ，そんなことをおっしゃるのは．ねえ」

日本語の母語話者であれば，まず間違いなく，最初の発話者は男性で，次の発話者は女性であると推測する．それが可能なのは，これらの発話文のなかに，発話者のある種の属性 (この場合は性別) を推測する手がかりが埋め込まれているから，と考えるのが自然である．

本論文では，発話者の属性から構成される，ある種の典型的な発話者のタイプを**キャラクタ**と呼び，そのようなキャラクタと強く結びつくことを**キャラクタ性を持つ**とよぶ．言い換えるならば，キャラクタ (キャラクタ性) とは，それぞれの個人がもつ個性をグループ化・典型化した概念である．そのようなキャラクタ性を持つ言語表現として，金水の役割語 [2] がよく知られている．

文にキャラクタ性を埋め込むことができるのは，ある内容を伝えるための表層文に自由度があるからである．たとえば，例文 (1) の最初の発話文と同じ内容を伝える発話文として，次のような文が考えられる．

(2) 「俺がだれだか，知っているんか」
(3) 「あたしがだれなのか，ご存知なのですか」

例文 (2) の場合は，例文 (1) より粗野なキャラクタが想像される．例文 (3) では，発話者は女性と推測される．これらの発話文を見比べると，一人称表現 (ぼく/ 俺/あたし)，疑問表現 (だれだか/だれなのか)，述語表現 (知っている/ご存知)，形式名詞 (の/ん) に自由度があり，可能な選択肢の中のどの表現を選択するかによって，読み手が想像するキャラクタが変わると考えられる．

表 1: 1 セットの発話文

キャラクタ	発話文
(元の発話文)	鹿の被害が報告されてますよね
女性-10 歳-親密	鹿の被害が報告されてるね
男性-10 歳-親密	鹿が出てきて困ったって話を聞くよね
女性-25 歳-親密	鹿の被害があるみたいね
女性-25 歳-非親密	鹿の被害が報告されてるんですって
男性-25 歳-親密	鹿の被害が報告されてるね
男性-25 歳-非親密	鹿の被害が報告されていますね
女性-50 歳-親密	鹿の被害がよく出てるね
男性-50 歳-親密	鹿の被害が報告されてるよね
女性-75 歳-親密	鹿の被害が報告されてるわね
男性-75 歳-親密	鹿の被害が報告されとるんじゃな

キャラクタ性を持った文を生成する文生成器を実現するためには，準備段階として，次のことが必要である．

1. 文のどこに (どのような構成要素に)，キャラクタ性を埋め込む自由度があるかを調べる
2. それらの要素それぞれに対して，選択肢を列挙する
3. それぞれの選択肢とキャラクタの関連性 (相関) を求める

これらのうち，本稿では，1 を明らかにするために行なった調査について報告する．

2 調査対象

本調査では，NTT メディアインテリジェンス研究所が作成した発話文データベース [3] を利用した．このデータベースは，対話システム用に人手で構築された発話データと，それを特定のキャラクタを強調するように人手で書き換えた発話データから構成される．特定のキャラクタとしては，性別，年齢，会話相手との親密度，の 3 属性で構成される 10 種類が設定されている．すなわち，元の 1 つの発話文に対して，書き換えられた 10 種類の発話文が付与されている．以下では，この計 11 種類の発話文を 1 セットと呼ぶ．データベー

表 2: 書き替え箇所の抽出に用いた表

(元の発話文)	鹿の被害が		報告される	テる			ます	よね
女性-10 歳-親密	鹿の被害が		報告される	テる				ね
男性-10 歳-親密	鹿が出てきて困ったって		話を聞く					よね
女性-25 歳-親密	鹿の被害が		ある		みたい			ね
女性-25 歳-非親密	鹿の被害が		報告される	テる		んです		って
男性-25 歳-親密	鹿の被害が		報告される	テる				ね
男性-25 歳-非親密	鹿の被害が		報告される	ている			ます	ね
女性-50 歳-親密	鹿の被害が	よく	出る	テる				ね
男性-50 歳-親密	鹿の被害が		報告される	テる				よね
女性-75 歳-親密	鹿の被害が		報告される	テる				わね
男性-75 歳-親密	鹿の被害が		報告される	とる		んじゃ		な

スは，平叙文 307 セットと疑問文 325 セットから構成されている．表 1 にデータベースの 1 セットを示す．

この表からわかるように，それぞれのキャラクタに対して，異なる書き換えが行われている[1]．すなわち，これらの書き換え部分は，自由度を持つ要素の候補であり，キャラクタ性を持つ表現の候補とみなすことができる．本研究では，この発話文データベースのうち，平叙文 100 セットを対象に，書き換え箇所を詳細に調査した．

3 書き換え箇所の調査

書き換え箇所の調査は，以下の手順で行なった．

1. それぞれのセットに対して，表 2 に示すような表を作成し，書き換えられている部分を抽出した．その際，複数の書き換えが関与していると考えられるものは，それらを分解することとした．たとえば，「報告されてますよね/報告されてるね」は，「ます/(削除)」，「よね/ね」の 2 つの書き換えに分割する．おおよそ，この表の縦 1 列が，書き換えの 1 単位を表す．
2. 抽出した書き換えを言語単位および品詞等の形式に基づいて整理した．この際，意味が変位していると考えられる書き換えは，キャラクタの違いだけを反映しているとは言えないため，除外した．

表 3 に，除外した書き換えの例を示す．この表に示すように，対話相手への敬意の変更を伴うもの，アスペクトの変更を伴うもの，モダリティの変更を伴うもの，その他の意味の変更，は，意味が変位している言い換えとして除外した．調査に使用した発話データベースの作成に当たっては，作業者に「元の発話の意味を維持」することを要請している [3] が，意味の維持と

[1]書き換えを行なった作業者は，それぞれのキャラクタで異なり，一人の作業者が 1 セットの書き換えを行なったわけでない．それにも関わらず，多くのセットにおいて，書き換え後の発話文は，キャラクタ毎に異なる．

表 3: 意味が変位する書き換えの例

1. 話し手の敬意の変更		
常体/敬体	あるんだね	あるんですね
丁寧表現	寿司屋	お寿司屋さん
	話を聞くと	話を伺うと
	休みます	休まさせていただきます
2. アスペクトの変更		
テいる	気になっています	気になります
テしまう	おどろきました	驚いちゃった
テくる	気がする	気がしてくる
テみる	なりたいです	なってみたいのう
3. モダリティの変更		
助動詞	すばらしいですね	すばらしいことだね
	異なるといえる	違うといえるでしょうね
	採っていますね	採っているようですね
	習慣にした方がいいですよ	習慣にした方が良いと思います
係助詞	苦手な人が多い	苦手な人は多い
	怪しくはない	怪しくない
	スマホさえあれば	スマホがあれば
4. その他の意味の変更		
上位/下位	生春巻きが嫌いな人	生春巻きが苦手な人
	大学時代に	学生のときに
明示/暗示	体調を崩した時に便利です	体調を崩した時に飲むと便利です
	寅さんの姿が見れなくなって	寅さんが見れなくなって
説明	鹿の被害が	鹿が出てきて困ったって
	先進国になると	国がお金持ちになると
	接客業は	人とお話をすることは

変位の境界は明確ではない．今回の調査では，意味の維持をかなり保守的に捉えた．

4 検討

抽出した書き換えを，内容表現と機能表現に分けて整理した結果を，それぞれ表 4 と表 5 に示す．これらの表に示すように，内容表現は，同一語彙の異形の交替と同義表現の交替に大きく分類した．機能表現は，同一語彙の異形の交替，同義表現の交替，構造の変更を伴う交替，挿入・削除，の 4 種類に大きく分類した．なお，終助詞の書き換え (交替，挿入・削除) は，他の文末要素 (主に判定詞) との関係を考える必要があるた

表 4: 内容表現の書き換え

1. 同一語彙の異形の交替		
名詞	マリオカート	マリカー (省略形)
形容詞	小さい	ちっちゃい
副詞	とても	とっても
2. 同義表現の交替		
名詞	トイレ	洗面所
副詞	比較的	だいたい
動詞	低下します	下がります
形容詞	おいしい	うまい
オノマトペ	しゅわしゅわして	しゅわっとして

め，これらの結果から除外してある．

これらの表から，次のことが観察される．

1. 文のいたるところに自由度があり，それらの書き換え例が実際に存在する．
2. 書き換えの言語単位は，語，文節だけでなく，より大きな単位も存在する．
3. 内容表現の書き換えは，同義表現への書き換えが主である．ほとんどの語には同義表現が存在するため，大きな自由度がある．しかし，その一方で，特定の語彙とキャラクタの間に強い関連性があるとは一概には言えない．
4. 機能表現の書き換えでは，同義表現への書き換えとともに，同一語彙の異形への書き換えが多数観察された．これは，対象とした文が発話文 (話し言葉) であることが大きく寄与していると考えられる[2]．
5. 機能表現の同義表現への書き換えは，推測・伝聞・時間・理由・条件などの機能に基づいて分類した．同一の機能を担う表現が複数存在するという事実が，この自由度をもたらしている．これらの表現の変更は，人間にとっては容易であり，かつ，文の意味を大きく変える危険性が低い．また，異形も多く存在する．そのため，キャラクタを反映させやすいと作業者が判断したと考えられる．
6. 構造の変更を伴う交替のうち，助詞の交替や語順の変更は，日本語文法がそのような自由度を持っているからである．複合語の展開は，構造的に展開されるものをここに分類したが，これは，説明的に展開されるもの (「接客業/人とお話しすること」) へと連続的につながっている．
7. 機能語の挿入・削除は，話し言葉に典型的に見られる現象である．これも，調査対象が発話文であることが大きく寄与している．

上記の観察結果から，キャラクタ性を持った文を生成する文生成器の実現に対して，次のような方針が得られる．

1. 文のいたるところに自由度があるので，これらの自由度をすべて実装するのはハードルが高い．キャラクタと強く結びついていると考えられる要素に重点を置く必要がある．それがどのような要素であるかは，もう少し調査を進める必要があるが，おそらく，終助詞を含む文末表現と，特定の機能を表す機能表現の 2 つが，有力な候補である．文節機能部の書き換えに焦点を当てたこれまでの実装 [4] は，正しい方向性を示している．
2. 文末表現を含む機能表現の同義表現を網羅的に列挙する必要がある．日本語機能表現辞書『つつじ』[5] を出発点としてこれを拡張し，内容語を含む表現も取り入れていく必要がある．
3. 文生成器 Haori[6, 7] は，文節単位の生成を行なうが，それでは不十分である．複数の文節の生成を制御できる仕組みと，それへの入力の記述形式を定める必要がある．

参考文献

[1] 星新一. ノックの音が. 新潮文庫, 1985.

[2] 金水敏（編）. 〈役割語〉小辞典. 研究社, 2014.

[3] 宮崎千明, 平野徹, 東中竜一郎, 牧野俊朗, 松尾義博, 佐藤理史. 話者のキャラクタ性に寄与する言語表現の基礎的分析. 言語処理学会 第 20 回年次大会 発表論文集, pp. 232–235, 2014.

[4] 宮崎千明, 平野徹, 東中竜一郎, 牧野俊朗, 松尾義博, 佐藤理史. 文節機能部の確率的書き換えによる言語表現のキャラクタ性変換. 人工知能学会論文誌, Vol. 31, No. 1, 2016.

[5] 松吉俊, 佐藤理史, 宇津呂武仁. 日本語機能表現辞書の編纂. 自然言語処理, Vol. 14, No. 5, pp. 123–146, 2007.

[6] 佐藤理史. 「文生成器を作る」とはどういうことか. 言語処理学会第 21 回年次大会発表論文集, pp. 1080–1083, 2015.

[7] 緒方健人, 佐藤理史, 松崎拓也. 文節木の段階的実体化による日本語文生成器の作成. 2015 年度人工知能学会全国大会論文集, 2015.

[2] これに対して，書き言葉では，表記に自由度がある．

表 5: 機能表現の書き換え

1. 同一語彙の異形の交替			
形式名詞	の/ん	できるのは	できるんは
動詞	て [い] る/とる/ておる	報告されてますよね	報告されていますね
		向いていると思います	向いとると思うんじゃ
		信頼しています	信頼しておるんじゃ
	てしまう/ちゃう	とってしまうんですよね	とっちゃうんですよね
接尾辞	ない/ん	出来ないといけない	出来んといかん
	[ら] れる	食べれたら	食べられたら
判定詞	だ/じゃ	NBA だ	NBA じゃ
活用語尾	-だ/じゃ	大変だった	大変じゃった
	[っ] て/うて	近くて	近くって
		近くて	近うて
助動詞	んだ/の	気になってるんだ	気になってるのよ
助詞	は/って (提題)	長友選手は	長友選手って
	と/って (引用)	異なるといえる	違うっていえる
	という/っていう/っちゅう (同格)	左ハンドルという点	左ハンドルっていう点
		左ハンドルという点	左ハンドルっちゅう点
	では/じゃ	嫌いではない	嫌いじゃない
2. 同義表現の交替			
推測表現	ようだ/らしい/みたいだ	出たみたいです	出たようです
		いいらしいです	いいみたい
伝聞表現	そうだ/らしい/みたいだ/んだって	欠かさないそうです	欠かさないんだって
		苦労するらしいですよ	大変なんだって
		出たみたいだよ	出たんだって
意志表現	予定だ/つもりだ	行く予定です	行くつもり
時間表現	で/の時に	高校くらいで	高校くらいの時に
理由表現	ので/から	変わりやすいから	変わりやすいので
条件表現	-れば/たら	あれば	あったら
	なら/のだったら	遊ぶなら	遊ぶんだったら
	と/-たら	聞くと	聞いたら
順接・逆接表現	んだけど/んじゃが	使ってるんだけど	使ってるんじゃが
	が/けれど/けど	使っていましたが	使っていたけれど
		出たことあるけど	出たことあるが
		聞いた事あるけど	聞いた事があるけれど
並列表現	など/とか	Facebook などの	Facebook とかの
	や/とか	バナナや	バナナとか
同格表現	の/という	ハチミツとクローバーの映画	ハチクロっていう映画
すべてを表す表現	どんな〜でも/全部	どんな動物でも可愛い	動物は全部可愛い
疑問表現	どう/どんなふうに	どうなるんでしょう	どんなふうになるのかしら
取り立て表現	でも/だって/であっても	どんな動物でも	どんな動物だって
		どんな動物でも	どんな動物であっても
	でも/だといっても	同じ関東地方でも	同じ関東地方じゃゆーても
	でも/にも	どこの水族館にも	どこの水族館でも
3. 構造の変更を伴う交替			
ガ/ヲ交替	を/が	生春巻きを苦手な人	生春巻きが苦手な人
ガ/ノ交替	の/が	フカヒレの入ったチャーハン	フカヒレが入ったチャーハン
ガ/ニ交替	が/に	医者と弁護士に知り合いがいる	お医者さんと弁護士さんが知り合いにいる
語順の変更		中国の今後の	これからの中国の
複合語の展開		トイレ文化	洗面所の文化
4. 挿入・削除			
形式名詞	の	見上げるには	見上げるのには
助詞	が (格)	想像つかないな	想像がつきません
	に (格)	子供の頃に	子供の頃,
	を (格)	いいところ知っていますよ	いいところを知っています
	は (係)	私は朝,	私は朝は,
	など (副)	内紛や契約問題で	内紛や契約問題などで
判定詞	だ	出たみたいだよ	出たみたいよ