TR-IT-0141

代名詞の先行詞推定に関する センタリング理論の評価

Evaluation of Centering Theory with regard to Pronoun Antecedent Determination

荒川　直哉
ARAKAWA, Naoya

1995年12月

概要

日本語談話中の代名詞的表現の先行詞を推定するシステムを作成し、先行詞候補の文法機能と先行詞としての優先度との関連を調査した。センタリング理論によると、代名詞に先行する名詞句はその表層格やトピックなどの文法機能によって先行詞候補としての優先度が異なると予測される。しかし今回の調査ではセンタリング理論での通説を裏付けるような優先度の違いは認められなかった。調査のために使用したコーパスは㈱ＡＴＲ音声翻訳通信研究所が作成した旅行関連の２２６会話である。

1．はじめに

代名詞あるいはそれに準ずる表現は先行する談話（またはテキスト）に現れる表現を先行詞とし、その先行詞が指示するものと同じものを指示することがある。以下では、代名詞的な表現が先行詞を持つ場合、代名詞的表現がその先行詞と*照応*するといい、代名詞的表現のこうした用法を*照応用法*と呼ぶ。通常、代名詞的表現の先行詞となりうる表現は数多く存在するから、話者がそのうちどれを実際の先行詞として用いているかを推定することが言語理解のために必要である。

実際の先行詞推定プロセスには、先行詞候補を絞り込む統語論的および意味論的制約や、先行詞としての優先度を決定する諸条件が関与すると考えられる。統語論的制約の例としては、通常の（再帰的[1]でない）代名詞は同じ動詞に直接支配される名詞句を先行詞として持たないという制約が挙げられる。意味論的制約としては、代名詞とその先行詞は意味属性（素性）が一致すべきだということが挙げられる（たとえば「彼女」は通常「太郎」を先行詞としない）。先行詞としての尤度を決定する要素としては、代名詞的表現と先行詞候補の間の距離や、先行詞候補の文法機能を考えることができる。今回の調査で作成したシステムも、この考えに沿い、代名詞的表現に先行する名詞のうちで統語論的制約を満たし、意味素性が代名詞的表現と一致するものに、代名詞的表現との距離、助詞の情報などから優先度を与え、最も優先度の高いものを先行詞第１候補とする、というプロセスを採用している。

今回の調査の目的は、代名詞的表現の先行詞推定プロセスにおけるセンタリング理論の検証である。センタリング理論とは以下に述べるとおり談話要素の顕著さに関する理論であり、談話要素の顕著さが先行詞推定に役立つと主張するものである。本調査ではこの主張が裏付けられるかどうかを検証する。

2．センタリング理論

「センタ」とは談話において、ある時点で「最も顕著（salient）な談話要素であり」（竹下, 1992）、代名詞やゼロ代名詞[2]の先行詞として優先的に参照されると考えられている。センタリング理論は２つの要素からなる。１つは、上で述べたように、先行詞候補は文中での機能により顕著さが異なるという仮説であり、もう１つは一度先行詞となった表現はその顕著さを持続して持つという仮説である。

2.1　文法的機能によるセンタのランキング

Walker ら (1990) によると、文中の照応対象候補はその文中の文法的な機能によって顕著さが異なる。彼女らの仮説によると、その順位は次のようになる。

トピック＞共感＞主語＞目的語＞その他

ここで、トピックは、日本語では係助詞「は」に支配される要素とみなされる。共感 (Kuno, 1976) とは動作などを言い表すときの視点であるが、日本語では例えば「くれる」「くる」などの助動詞的用法に付随して現われてくる。

2.2　顕著さの持続

Joshi ら(1981)は、Cb (Backward-looking Center) および Cf (Forward-looking Center(s)) という２種類の談話のセンタを考案した。Cb は談話のある時点で最も顕著な要素である。Cb は代

---

[1] reflexive

[2] 文中で省略されている(代)名詞的要素

名詞の先行詞としての優先度が高いと考えられるが、代名詞の先行詞となることによって（後ろ向きに）発見されることからこの名前がつけられたと考えられる。Cb には変化を嫌う性質があるものと考えられている。Allen (1987) のフォーカス仮説によれば「代名詞先行詞候補の中には第１候補となるものが存在し、それは必要がなければ変化しない」(p. 408) ということである。

Cf はある時点で、代名詞の先行詞となりうる談話要素のリストである。Cf には 2.1 で述べたような文法機能によるランキングが与えられると考えられる。

Takada & Doi (1994) は、一度先行詞となった要素は、Cb であるかとうかにかかわらず、そうでない要素より顕著であると考え、センタリング理論を拡張している。

３．研究対象

上で述べたように、本調査の目的はセンタリング理論の検証である。検証において用いるコーパスは日本語の旅行に関する２２６会話である。これはＡＴＲ談話コーパスと呼ばれているものの一部である[3]。談話の多くは電話を介した会話であるが窓口業務の会話も含んでいる。

・対象とする照応

調査対象となる照応は、代名詞（あるいはそれに準ずる表現）が先行する名詞句を先行詞とする場合である。日本語の場合、何をもって代名詞とするか（あるいは代名詞そのものが存在するか）については定説がないが、今回の実験では調査対象とする照応用法を持つ代名詞的表現として以下のものを定めた。

「そちら、あちら、それ、これ、それら、これら、彼、彼女、そこ、あそこ、それぞれ、どちら、そのとき、その時」

ここで「どちら」および「それぞれ」は並列表現のみを先行詞候補とする。

４．先行詞推定プログラム

今回の調査のために先行詞推定プログラムを作成した。処理の概要は以下のとおりである。

4.1 処理概要

照応用法を持つ代名詞的表現に先行する名詞のうち、意味素性が代名詞的表現と一致し（意味論的制約）、代名詞的表現と互いに同一の動詞（名詞）の格要素（修飾要素）でない（統語論的制約）ものについて、文法機能および語間距離に応じたスコアリングを施し、最もスコアの高いものを照応対象として選ぶ。

代名詞的表現の語彙的な意味素性と名詞句の意味素性のマッチングを行い、単一化できない場合は不適格とする（意味論的な適格性の検査）。また、同一の動詞の格要素か修飾要素であるような名詞句は不適格とする（統語論的な適格性の検査）。

文法機能に基づくスコアリングでは、

1) 照応対象候補が助詞「は」に支配されているか (TOPIC)
2) 主文の主語（格助詞「が」に直接支配されるもの）または主語と同格の補語か (NOM)
3) 主文の目的語（格助詞「を」または「に」の目的語）か (OBJ)
4) 主文の主語でもなく目的語でもないが、主動詞を直接修飾する要素か (MOD)
5) それ以外か (DEFAULT)

によって、それぞれ別のスコアを与えた。（「共感」によるスコアリングは行わなかった。）

---

[3] 実験時点において人手による格構造の解析が終了している会話をすべて使用した。

さらに、直前に第1先行詞候補として選ばれた語に特別のマーク (D-FOCUS [談話フォーカス]) を与え、高スコアを与えるようにした。この「談話フォーカス」は、ある文で唯一の代名詞的表現の先行詞となるか、相続く2つ以上の文で継続して代名詞的表現の唯一の先行詞に選ばれることによって認識される。D-FOCUS が与えられた表現は、上で説明した Joshi らの Cb に相当すると考えられる。また、Takada & Doi (1994) に従い、過去に第1先行詞候補として選ばれたすべての語に別のマーク (PRO-SALIENT) を与え、スコアを追加するようにした。なお、日本語においてはゼロ代名詞（省略された代名詞）の先行詞として選ばれる表現も上記のマークを与える対象として考慮にいれるべきであるが、今回の調査対象には入れていない。そのため、本調査では先行詞候補への D-FOCUS, PRO-SALIENT のマーキングは不完全なものである。

より近い語に優先度を与えるべく、スコアには代名詞的表現と先行詞候補間の形態素の数に対し、一定の減衰率 (0.98) を設定した。なお、個々の句読点も形態素の1つとみなした。

以上をまとめると次のようになる。各代名詞的表現に対し先行する名詞に先行詞候補として下記のスコアが与えられ、スコアの最も高いものが先行詞第1候補として選ばれる。

$$\text{スコア} = \text{decl\_rate}^{\text{utt\_dist}} \times (\text{centr\_score} + \text{focus\_score}) \times \text{syn\_score} \times \text{sem\_score}$$

ここで、

decl_rate は減衰率で、utt_dist は代名詞的表現と名詞の間の形態素の数である。

centr_score は文法機能別のスコア倍率である。

focus_score は談話フォーカス関連で、D-FOCUS の追加スコアあるいは PRO-SALIENT の追加スコアのうち大きいほうを与える。

syn_score は名詞句が代名詞的表現と同じ動詞または名詞に掛かる要素である場合に0となり、その他の場合は1である。

sem_score は名詞と代名詞的表現の意味素性が単一化できないときのみ0となり、その他の場合は1である。

4.2 使用したデータ

入力として談話の格構造解析結果を用いた。この格構造は、人間の修正が加わったもので、ほぼ正しいと考えられる。格構造には先行関係を明らかにするため、また、代名詞的表現と先行詞候補間の距離による顕著さの減衰を計算するために形態素の出現順に番号を振った。

名詞と代名詞的表現に対しては意味素性情報を記述した辞書を用意し、格構造に意味素性を埋め込んで利用した。意味素性としては、 LOC（場所)、 NUM（数)、HUM（人間)、 GENDER（性)、TLOC（時間的位置）を用いた（多義性などにより不定の場合は素性の指定を行わない）。

5．実験

２２６会話中の照応用法を持つ代名詞的表現のうち、会話中に名詞の先行詞を持つ表現について正解の先行詞を記述したファイルを作成しておき、先行詞推定プログラムの出力とマッチングを行った。スコアリングのパラメータをいろいろ変化させて正解率を求めた。

5.1 結果概要

1. 現在得られている実験結果からは、センタリング理論が先行詞候補推定に「非常に」有用であるということはいえない。すなわち、センタリング理論に沿って、助詞「は」に支配される語や主語の先行詞候補としての優先順位を上げても、先行詞の推定率は上がらない。

2. 代名詞的表現と先行詞の平均距離を先行詞の文法機能別に調べた。より顕著な文法機能をもつ先行詞はより遠くから照応することができると考えると、代名詞的表現先行詞間の平均距離はセンタリング順位を反映していると考えることもできる。この考えが正しければ文法機能別のセンタリング順位は

   SUBJ [4]> TOPIC > DEFAULT > MOD > OBJ

   となり、通説の

   TOPIC > SUBJ > OBJ > MOD > DEFAULT

   とは異なる。
3. D-FOCUS,PRO-SA:LIENT のマーキングに基づくスコアリングを用いた場合、若干正解率の上昇を見たが、やはり「非常に」有用であるということはいえない。

5.2 先行詞候補の文法機能に対するスコア倍率[5]と成績[6]

倍率	1.0	1.04	1.08	1.12	1.16	1.20	1.40
TOPIC	151	151	150	149	148	145	140
NOM	151	151	150	152	152	152	144
OBJ	151	151	151	151	150	149	143
MOD	151	153	152	150	151	149	144
Gradation	151	150	149	149	145	143	130

ここで倍率 1.0 の列は文法機能によるスコア変化を行わない対照である（成績：１５１）。TOPIC、NOM、OBJ、MOD の行は、それらの要素のみスコアを増加させ、他の要素については増加分を１／４したものを減じた場合である（例えば、TOPIC の倍率が 1.04 の場合、NOM, OBJ, MOD, DEFAULT の倍率は 0.99 である）。最後の「Gradation」では、倍率 1+X を TOPIC に、1+X／2 を NOM に、1 を OBJ に、1−X／2 を MOD に、1−X を DEFAULT に、それぞれ与えた。例えば増加率 1.08 の場合、TOPIC に倍率 1.08、 NOM に 1.04、 OBJ に 1.0、MOD に 9.96、DEFAULT に 9.92 が、それぞれ与えられることになる。

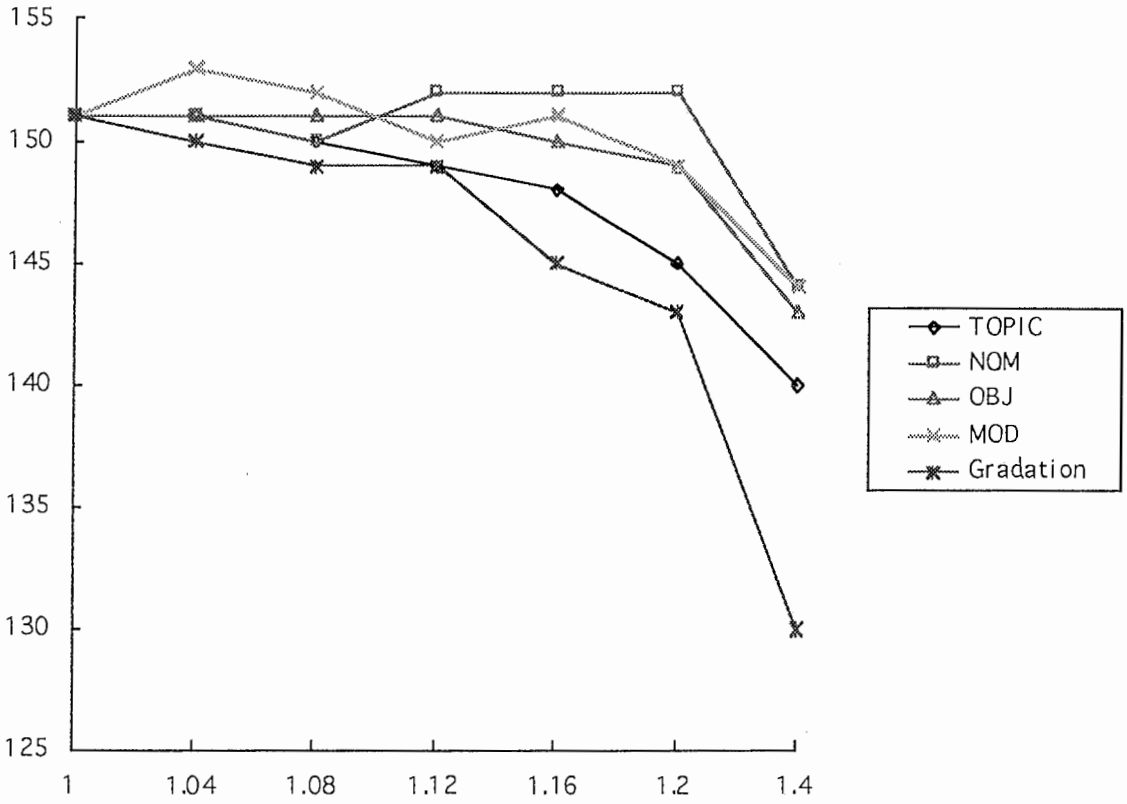
全体として、特定要素へのスコアを上げるにつれ成績が悪化するということが言える（次頁グラフ参照）。すなわち、最も近い先行詞候補を優先するのが一般的によいということである。「Gradation」の成績悪化率が大きいのは、大多数を占める DEFAULT に低い率 (1−X) が与えられ、最も近い先行詞候補が選ばれにくいためと考えられる。

---

[4] 格助詞「が」に直接支配される主語

[5] 4.1 の式の centr_score.

[6] 以下のすべての実験において形態素当たりの減衰率は 0.98 に固定してある。実験パラメータはすべてこの減衰率に対し相対的なものであって、絶対値は意味を持たない。



先行詞候補が NOM の場合、倍率が 1.2 を超えて良好な成績を保っているが、これはここでの検証とは無関係な原因によるものである。業務対話では「（こちらは）N でございます。」などという形で会話が始まり、客側が相手のいる場所を「そちら」という代名詞で指すことが多い。こうした場合（N は主語と同格の名詞として NOM 扱いされるのであるが）、代名詞と先行詞の距離はかなり長いものとなりうる。本来の主語（SUBJ：格助詞「が」に直接支配されるもの）のみにスコアを追加した場合の統計を以下に示す。

倍率	1.0	1.04	1.08	1.12	1.16	1.20	1.40
SUBJ	151	151	150	148	148	147	141

さらに「そちら」を実験対象から除いた場合の統計を以下に示す。

倍率	1.0	1.04	1.08	1.12	1.16	1.20	1.40
NOM	129	129	128	127	127	126	118
SUBJ	129	129	128	126	126	125	120

SUBJ のみにスコアを追加した場合、あるいは「そちら」を取り除いた場合、NOM または SUBJ へのスコアリング率の増加に対し、成績は単調減少する。

5.3 代名詞的表現と先行詞の間の距離

正解ファイルを分析して代名詞的表現と先行詞の平均距離を先行詞の文法機能別に調べた結果を以下の表に示す。文法機能 IDEN は名詞が述語的に「A です」などの形で用いられていることを示す[7]。

[7] 分布については付録を参照のこと

文法機能	総数	平均発話距離	平均語間距離
TOPIC	48	1.81	21.1
SUBJ	17	2.65	30.8
IDEN	59	7.64	85.9
OBJ	6	1.00	12.3
MOD	31	1.13	13.6
DEFAULT	111	1.60	21.6
全体	272	2.95	34.9

注）２２６対話の１発話あたりの平均形態素数は 11.2 個である。

上の表で、IDEN すなわち述語として現われる名詞を先行詞とする場合の平均発話距離と平均語間距離が他の文法機能に対するものよりとびぬけて大きい理由は、5.2 で述べたように、挨拶の発話にあらわれる場所名を後に「そちら」で照応する場合が多いからである。「そちら」を除いた統計を以下に示す。

文法機能	総数	平均発話距離	平均語間距離
TOPIC	39	1.74	19.3
SUBJ	16	2.75	32.3
IDEN	29	1.79	13.6
OBJ	5	1.00	12.2
MOD	26	1.19	13.9
DEFAULT	98	1.32	19.1
全体	213	1.55	18.6

ここで先行詞が SUBJ の場合に平均語間距離が比較的大きいことに注意されたい。もし平均語間距離が、先行詞候補の文法機能による顕著さの程度を示すのなら、その度合は、

SUBJ>IDEN>TOPIC>DEFAULT>MOD>OBJ

の順となる。

5.4 D-FOCUS, PRO-SALIENT の影響

D-FOCUS に与えるスコア増分に対する成績の表

D-FOCUS	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
成績	151	151	151	153	151	151	151	148	146

PRO-SALIENT に与えるスコア増分に対する成績の表

PRO-SALIENT	0	0.1	0.2	0.3	0.4	0.5	143	133
成績	151	151	149	150	149	144	142	133

以上のデータは文法機能別にスコアの差を与えていない場合の成績である。D-FOCUS に 0.4 のスコア増を与えた場合、若干先行詞推定の成績がよくなっているが、これが偶然によるものかどうかは、今回の実験ではデータ不足のため不明である。D-FOCUS, PRO-SALIENT マーキングの利用に関しては、今回の実験は以下の理由により予備実験の域を出ない。

1) ゼロ代名詞を考慮していない。

2) D-FOCUS に関しては、コーパス中で代名詞的表現が近接して多数出現するということが稀なので、データ量が十分ではない。
3) 先行詞推定率が低いので、誤った D-FOCUS, PRO-SALIENT のマーキングをしてしまうことが多い。

5.5 正解率

今回の実験の目的はセンタリング理論の検証であり、正解率は実験を行うために最低限のデータ件数を保証することができる程度に高ければよい。照応用法を持つ代名詞的表現は、２２６対話中４３３個出現した。そのうち対話中の名詞を先行詞として持つものは２７２個である。現在のところ先行詞推定システムの最良の成績ではそのうち１５３個を正しく選んでいるから、最良の正解率は約５６％である。先行詞推定システムは、実用を目標とするならば先行詞が名詞かどうかを判別しなければならないが、その点については今回の実験ではまったく考慮していない。

６．調査の発展

この調査の自然な発展として、ゼロ代名詞を加えた場合を考えることができる。特に、 D-FOCUS, PRO-SALIENT のマーキングが（ゼロ）代名詞の先行詞推定に与える影響に関してより確かな結果を得るためには、ゼロ代名詞も含めた代名詞的表現全体の先行詞推定について検討する必要がある。

代名詞的表現の照応に関して調査可能なもう１つの分野は、談話構造が先行詞推定に与える影響である。談話中には質問-応答パターンなどの談話構造があるが、こうした構造に対して距離を導入し、本実験で用いた形態素数による距離と、先行詞推定に関して比較検討することが考えられる。

文献

Allen, J.: *Natural Language Understanding*, Chapt. 14. The Benjamin/Cummings Pub. Co. (1987).

Kuno, S.: "Subject, theme and speaker's empathy: A reexamination of relativization phenomena." in *Subject and Topic*, C. Li ed., pp. 417-444. Academic Press (1976).

Takada, S. and Doi, N.: Centering in Japanese: A Step Towards Better Interpretation of Pronouns and Zero-Pronouns, COLING 94 Proceedings, pp. 1151-6 (1994).

竹下敦：「文タイプ情報を用いた話題構造の認識」、人工知能学会研究会資料, SIG-SLUD-9203-2 (1992).

Walker, M., Iida, M. and Cote, S.: Japanese Discourse and the Process of Centering, *Computational Linguistics*, Vol. 20 #2, pp. 193-231 (1994).

付　録

各代名詞的表現の用法

今回調査の対象となった代名詞（およびそれに準ずる表現）に関して対象となった２２６対話中の照応、非照応用法の数を調べた結果を以下に示す。「正解」とあるのは、照応用法の代名詞（およびそれに準ずる表現）のうち、実験に使用した先行詞推定プログラムが正しく推定した先行詞の数である。

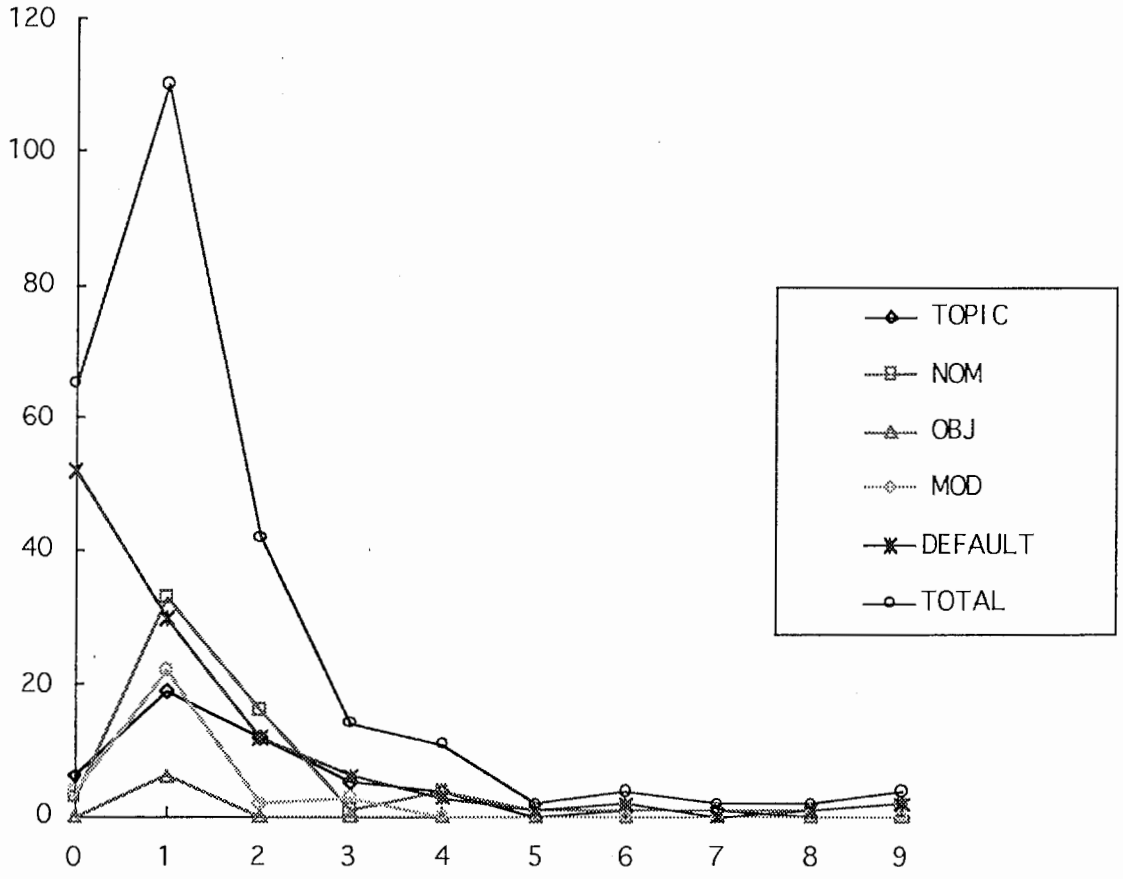
	総数	非照応用法	照応用法	正解	正解率
そちら	68	14	54	23	0.43
あちら	0	0	0	0	
それ	164	94	70	42	0.7
これ	51	28	23	18	0.78
それら	0	0	0	0	
これら	0	0	0	0	
彼	1	0	1	1	1
彼女	0	0	0	0	
あそこ	3	2	1	0	0.0
そこ	42	3	39	27	0.69
それぞれ	28	0	28	7	0.25
どちら	70	17	53	32	0.6
そのとき（時）	5	2	3	3	1
合計	432	160	272	153	0.56

非照応用法の内訳

・そちら	相手のいる場所	１１
	事態	2
・それ	事態、内容など	９０
	定型表現	2
	その他	2
・これ	直示用法	9
	事態、内容など	１５
	その他	4
・あそこ	直示用法	1
	その他	1
・そこ	直示用法	1
	事態、内容など	1
	その他	1
・どちら	疑問詞（場所／方角）	１２
	選択	4
	定型的表現	1
・そのとき	文で表現された状況	2

代名詞的表現と先行詞の距離の分布

文距離	0	1	2	3	4	5	6	7	8	9	10以上	全体
TOPIC	6	19	12	5	4	0	1	1	0	0	0	48
NOM	3	33	16	1	4	1	1	1	1	2	13	76
OBJ	0	6	0	0	0	0	0	0	0	0	0	6
MOD	4	22	2	3	0	0	0	0	0	0	0	31
DEFAULT	52	30	12	6	3	1	2	0	1	2	2	111
TOTAL	65	110	42	15	11	2	4	2	2	4	15	272



なお、最も代名詞的表現から遠い先行詞は４７文離れたところにあったが、代名詞は「そちら」で先行詞は対話のはじめに現われた「（こちらは）～です。」の「～」の部分であった。また、並列表現を先行詞に取る代名詞「それぞれ」「どちら」が比較的遠い先行詞と照応することも見られた。これは並列表現が比較的稀で、しかも明確な識別特徴をもっていることによるのかもしれない。