

省略補完における談話的制約の有効性の検証

藤澤 伸二 *

増山 繁 *

内藤 昭三 **

* 豊橋技術科学大学知識情報工学系

** NTT 基礎研究所

概 要

本稿では、先に提唱した、省略に関する4つの制約が実際の科学技術解説文(日経サイエンスの記事)と論説文(「朝日新聞」の天声人語)において省略語の復元を行なう上でどの程度有効であるかを検証するため、語の素性の一致制約に、さらにこれら4つの制約を加えれば、どの程度省略語補完の結果が改善されるかを実際に計算機実験により調べ、その結果について考察する。さらに、省略補完の精度を上げるため話題という概念を用いた。本稿では、この話題の推定を、2つの方法で行ない、実際の文章で調査し、その結果と考察を併せて述べる。

An Inspection on Effect of Discourse Constraints pertaining to Ellipsis Supplement

Shinji FUJISAWA*

Shigeru MASUYAMA*

Shozo NAITO**

*Dept. of Knowledge-based Info. Eng., Toyohashi Univ. of Tech.

**NTT Basic Research Laboratories

abstract

In this paper we inspect the effect of four discourse constraints pertaining to ellipsis in articles and essays when we supply ellipsis. We investigate, using computer experiments, how ellipsis supplement is improved by adding these constraints in addition to the feature agreement of words. We also use the idea of topic so that ellipsis supplement is improved. We estimate a topic by two methods, and we investigate it against articles and essays.

1 はじめに

本稿では、先に提唱した[1]、省略に関する4つの制約が、省略語の復元を行なう上でどの程度有効であるかを検証するため、語の素性の一致制約に、さらにこれら4つの制約を加えれば、どの程度省略語補完の結果が改善されるかの計算機実験を行ない、その結果について考察する。さらに、省略補完の精度を上げるため話題という概念を用いた。本稿では、この話題を推定する2つの方法を提案し、その有効性を実際の文章で検証し、その結果に対する考察を併せて述べる。なお、ここでは、省略をIPAL辞書[3]記述中の必須格要素の省略と定義する。また、検証テキストとしては、科学技術解説文(日経サイエンスの記事)と論説文(「朝日新聞」の天声人語)を用いた。

本稿の構成は、まず、2章で省略に関する4つの談話的制約を紹介し、3章で、語の素性の一致制約にこれら4つの制約を加えれば、どの程度省略語補完の結果が改善されるかを実際に計算機実験により調べた結果を述べる。4章では、省略語の先行詞の候補の中に正しい先行詞の含まれる割合を高くするための方法として、「話題」を用い、実際の文章(日経サイエンスと天声人語)で検証した結果を述べる。5章では、先の実験と調査から得られた考察を、また、6章では、今後の課題を述べる。

2 省略に関する談話的制約

文献[1]で提唱し、検証した照応・省略に関する8つの仮説(制約)のうち、省略に関する4つの制約は以下の通りである。なお、単文、複文、重文、最上位の節の定義は[1]に従う。

(制約1) 省略は段落間を跨らない。すなわち、省略の先行詞は、省略語と同一段落中にある。

(制約2) 省略は交差しない。すなわち、省略と先行詞の組の間には、交差が生じない。

(制約3) 最上位の節を除く各節において、省略の先行詞はその上位の節に含まれる。

(制約4) 単文中の省略、および、重文の先頭節で、しかも埋め込み節でない節中の省略に対する先行詞は、それより前の文中にある。

3 計算機による省略補完

3.1 補完方法

動詞・形容詞辞書としてIPAL動詞辞書、形容詞辞書[3]を、また、名詞辞書として角川類語新辞典[4]を用い、日経サイエンスの記事5編のそれぞれに対し、最初の7段落と「朝日新聞」の天声人語5編を対象として、省略語の補完を行なった。

日経サイエンスは、科学技術解説文であり、また、「朝日新聞」の天声人語は論説文であるが、日本語の文章の構造原理が良く現れ、しかも、充分推敲されている文章として、今回の実験にこれらの文章を使用した。

省略補完の方法は以下の通りである。まず最初に、IPAL動詞(形容詞)辞書に、ひとつの動詞(形容詞)に対して複数の文型がある場合、その中から現在処理している動詞(形容詞)と同一の文型を選択する。この文型の選択は、人手を介さず、計算機で行なっており、必須格形式が文章中の格形式と一致し、かつ、それらの格形式を伴う語のとり得る素性が一致する数が最大のものを選択している。但し、それらの一致する数が等しい場合は、一方でも一致しない数の最小のものを選び、それでも文型をひとつに絞り込めない場合は、IPAL辞書記述中の最初の文型を選択する。次に、その文型に記載されている必須格で現在処理している文にはないものを省略語とし、省略位置より前に出現した語の中からその必須格の持つ素性と同一の素性を持つ語を省略語の先行詞の候補とする。その上で、これらの候補に4つの制約を適用する。

3.2 実行結果

結果を表1から表6に示す。その中で、表1から表3は「日経サイエンス」、表4から表6は「天声人語」を対象に行なった結果である。

表1、表4は、すべての省略語の先行詞の候補のうち、それぞれの制約を満たすものの重複を含めた出現回数を示したものである。「制約なし」は、必須格の持つ素性と同一の素性を持ち、省略語より前に出現する語の総数である。表1、表4中の「未登録語」とは、今回使用した角川類語新辞典[4]に記載されていない語(例えば、シリコン、ネットワークなど)を先行詞とする省略語である。また、「省略語(b)」とは、「わたし(筆者、話者)は」や、「あなたがた(読み手、聴者)は」を指す省略語であり、その先行詞は、ほとんど文章に現れない。以下、「省略語(a)」とは、「省略語(b)」、「未登録語」を除く省略語を指す。

それぞれの出現回数は、日経サイエンスでは、「省略語(a)」が56回、「未登録語」が5回、「省略語(b)」が68回、天声人語では、「省略語(a)」が69回、「未登録語」が1回、「省略語(b)」が39回であった。

表2、表5は、「省略語(a)」の全出現回数(日経サイエンスの記事で56回、天声人語で69回)における省略語先行詞の候補の中に正しい先行詞が含まれている回数の割合である。先行詞が正しいかどうかの判定は人手によって行なった。

表3、表6は、「制約なし」時の候補数に対する制約使用時の候補数の割合を表したものである。制約を用いることによる候補数削減効果を示している。

	省略語(a)	未登録語	省略語(b)
制約なし	4211	568	3969
制約1	683	51	637
制約2	238	23	251
制約3	2585	436	3731
制約4	4211	568	3969
制約1,2	217	19	159
制約1,3	592	50	460
制約1,4	683	51	630
制約2,3	227	22	221
制約2,4	238	23	249
制約3,4	2838	436	3759
全制約	207	18	156

表1. 先行詞の候補の総数(語)
(日経サイエンス)

制約なし	86
制約1	80
制約2	43
制約3	77
制約4	86
制約1,2	50
制約1,3	75
制約1,4	80
制約2,3	43
制約2,4	39
制約3,4	73
全制約	45

表2. 省略語(a)の先行詞選択の正解率(%)
(日経サイエンス)

	省略語 (a)	未登録語	省略語 (b)
制約 1	16	9	16
制約 2	6	4	6
制約 3	61	77	94
制約 4	100	100	100
制約 1,2	5	3	4
制約 1,3	14	9	12
制約 1,4	16	9	16
制約 2,3	5	4	6
制約 2,4	6	4	6
制約 3,4	67	77	95
全制約	5	3	4

表 3. 「制約なし」に対する
候補数の割合 (%)
(日経サイエンス)

制約なし	86
制約 1	74
制約 2	35
制約 3	68
制約 4	86
制約 1,2	32
制約 1,3	61
制約 1,4	74
制約 2,3	39
制約 2,4	33
制約 3,4	68
全制約	39

表 5. 省略語 (a) の先行詞選択の正解率 (%)
(天声人語)

	省略語 (a)	未登録語	省略語 (b)
制約なし	3328	62	1643
制約 1	657	13	286
制約 2	184	1	111
制約 3	1806	2	1456
制約 4	3328	62	1643
制約 1,2	161	1	98
制約 1,3	421	2	221
制約 1,4	657	13	281
制約 2,3	183	2	95
制約 2,4	184	1	112
制約 3,4	1806	2	1456
全制約	167	2	66

表 4. 先行詞の候補の総数 (語)
(天声人語)

	省略語 (a)	未登録語	省略語 (b)
制約 1	20	21	17
制約 2	6	2	7
制約 3	54	3	89
制約 4	100	100	100
制約 1,2	5	2	6
制約 1,3	13	3	13
制約 1,4	20	21	17
制約 2,3	5	3	6
制約 2,4	6	2	7
制約 3,4	54	3	89
全制約	5	3	4

表 6. 「制約なし」に対する
候補数の割合 (%)
(天声人語)

4 省略補完の改善

4.1 話題の導入

省略は、聞き手が容易に復元できる情報を省く言語現象である [2]。聞き手が容易に復元できる情報とは、その文章の中で述べられてきたこと (あるいはそれから容易に推論されること) であり、古い、既知の情報である。一般に、話題と主題は同義に用いられることが多く [8]、異なった意味で用いられること [9] はあまりない。本稿では、省略され得る、この古い、既知の情報を話題と定義する。

日本語における主題に関してさまざまな研究がなされており、一般に文頭にある名詞句で助詞の「は」を伴っているものが主題と呼ばれてきた。これに対し、主題名詞句の談話レベルの研究により、これまでの文レベルにおける研究の不完全性が明らかになってきた [6]。特に、「は」で提示される名詞句が、ある談話情報の下では、新しい情報も表示可能であること [5]、さらに、古い情報が「名詞句 + は」形では表現されない時の条件 [7] 等の新しい結果が次々に発表された。これまでも「は」の機能に関する研究は多いが、共通している見解は、談話レベルを考慮に入れない限り、「は」の機能、さらには、主題一般の理解は不可能であるということである。

先の話題の定義により、省略語の先行詞は何かという問題を、その話題は何かという問題に置き換えることができる。現時点で主題や話題を確実に決定することは不可能なので、本稿では、以下の2つの方法で話題の推定を試みた。1番目の方法は、文章中の語に対応する分類番号 (角川類語新辞典 [4] に記載されている番号) の出現頻度の多い順からいくつかを選択し、話題とするという方法である。2番目の方法は、各文の主格 (「は」を伴う名詞句、「が」を伴う名詞句) を話題とするという方法で、主題より広い名詞句を先行詞の候補として残すことができる。

4.2 調査

調査は、前述した計算機による省略補完の実験で用いた、日経サイエンスの記事5編のそれぞれに対し、最初の7段落と天声人語5編を対象に行なった。

結果を表7から表10に示す。表7、表8は日経サイエンス、表9、表10は天声人語を調査対象とした。表7、表9は、文章の先頭から省略語までの間に存在する語の持つ分類番号を出現頻度の多い順から3番目までの番号 (多い順から3番目までの番号に、同頻度の番号が複数存在する場合は、それらをすべて含める) を選び、それらの分類番号と人間が正しいと判断した先行詞の持つ分類番号が小分類、あるいは、中分類で一致する割合を表したものである。それらは、素性の一致制約では正しい先行詞が選択できる場合 (日経サイエンスの記事で48回、天声人語で59回) と選択できない場合 (日経サイエンスの記事で8回、天声人語で10回) それぞれの割合と、さらに選択できる場合とできない場合とを合わせた全体 (日経サイエンスの記事で56回、天声人語で69回) の割合である。

表8、表10は、先行詞が、文章の先頭から省略語の間に存在する主格である割合を表したものである。これらの表も、素性の一致制約では正しい先行詞が選択できる場合の割合と選択できない場合の割合、さらに選択できる場合とできない場合とを合わせた全体の割合を表している。

		素性制約 を充足	素性制約 を不充足	全体
中 分 類	高頻度の語と 一致	56	63	57
	不一致	44	37	43
小 分 類	高頻度の語と 一致	46	63	48
	不一致	54	37	52

表7. 高頻度の語と先行詞の分類番号が一致する割合 (%)
(日経サイエンス)

	素性制約 を充足	素性制約 を不充足	全体
以前の主格と 一致	71	100	75
不一致	29	0	25

表 8. 先行詞が以前出現した
主格である割合 (%)
(日経サイエンス)

		素性制約 を充足	素性制約 を不充足	全体
中 分 類	高頻度の語と 一致	47	50	48
	不一致	53	50	52
小 分 類	高頻度の語と 一致	49	50	49
	不一致	51	50	51

表 9. 高頻度の語と先行詞の分類番号が
一致する割合 (%)
(天声人語)

	素性制約 を充足	素性制約 を不充足	全体
以前の主格と 一致	47	30	45
不一致	53	70	55

表 10. 先行詞が以前出現した
主格である割合 (%)
(天声人語)

5 考察

以上の結果から次のことが分かった。

談話的制約の有効性の検証について

- 制約 1、2 はすべての文を対象とし、しかも、省略語の先行詞の探索範囲を限定している制約なので、先行詞の候補をかなり削減できる。制約 3、ない

し、制約 4 は、制約 1、あるいは、制約 2 と併用する方が候補数の削減効果は大きい(表 1、3、4、6 参照)。

- 制約 3、4 が適用可能な単文、複文、重文の出現頻度は、実際の文章では少ない。また、制約 3 は先行詞の候補の探索範囲が現在処理中の文のみなので、探索範囲内に先行詞の候補となりえる語の存在する割合が低くなり、正解率(先行詞の候補の中に、人間が正しいと判断した先行詞の含まれる比率)が「制約なし」の場合に比べて低くなる。制約 4 の適用範囲は現在処理中の文のみなので、先行詞候補の削減効果は特に小さく、先行詞の候補数は、制約不使用時とほとんど変わらない(表 1、3、4、6 参照)。

- 制約 2 を用いて省略語の先行詞の候補を絞り込んだ場合、残された候補の中に、省略語の正しい先行詞が含まれる比率が低下する。その理由は以下の通りである。制約 2 を単独で使用すると、他の制約を併用した場合に比べ、先行詞の探索範囲が広くなり、候補として多くの先行詞以外の語を選んでしまう。ところが、現在の制約の使い方では、それらの語はすべて省略の非交差性により、後に出現する省略語に対する先行詞の候補からは除外されてしまうので、候補の中に正しい先行詞の含まれる比率が低下することになる(表 2、5 参照)。

- 制約 2 は、制約 1 と組み合わせて使用の方が効果的である。つまり、制約 1 で、ある程度省略語の先行詞の探索範囲を限定することで、省略の非交差性によって候補から正しい先行詞が除外される可能性を減少させることができる(表 2、5 参照)。

- 全制約を使用するより、制約 2 以外の制約を単独に、あるいは、いくつか組み合わせて使用した方が効果的である。この理由としては、すべての制約を用いたために、正しい先行詞が除外される可能性が高まるためであると

考えられる。制約 2 の単独使用、または、制約 2 との併用を除外する理由としては、制約 2 を使用した正解率が全制約を使用した正解率より低いからである (表 2、5 参照)。

- 省略語の先行詞の候補の中に正しい先行詞の含まれる割合は、全体的に見て、天声人語より日経サイエンスの方が良い。この結果は先の調査結果 [1] の裏付けになっている (表 1～6 参照)。
- 省略語の先行詞が素性一致制約のみの使用でも選択できない場合がある。その理由を例文とともに以下に示す (表 2、5 参照)。

1. 動詞・形容詞の文型選択の誤り

IPAL 辞書に記載されている文型が、ひとつの動詞・形容詞についても複数書かれている。現在は、その文型の選択を計算機で行なっているので、文型の選択誤りが存在する。

(例) ϕ (革命が) 広い層の民衆デモに発展する。(天声人語)

ϕ — 〈 HUM/ANI 〉

革命 — 〈 PHE/ABS 〉

〈 HUM 〉: 人間 (human)

〈 ANI 〉: 動物 (animal)

〈 PHE 〉: 現象名詞 (phenomenon)

〈 ABS 〉: 抽象名詞 (abstract)

2. 動詞・形容詞の文型に素性が未登録

IPAL 辞書に記載されている文型に、格は一致するが素性が一致しない場合、あるいは、素性は一致しても格が一致しない場合がある。つまり、IPAL 辞書の文型に素性(あるいは、文型そのもの)が未登録である。

(例) 疑問が ϕ (漁民から) 起こる。(天声人語)

ϕ — 〈 MEN 〉

漁民 — 〈 HUM 〉

〈 MEN 〉: 精神 (mental)

〈 HUM 〉: 人間 (human)

話題の導入について

- 処理している文章において (本稿での実験で用いた方法の意味で) 高頻度で使用されている語の分類番号と省略語の先行詞 (話題) の分類番号とが一致する割合は約 50 % であり、日経サイエンスの記事と天声人語、分類番号の中分類と小分類の間に大差は見られない。省略補完の際、省略語の先行詞の候補の中に正しい先行詞が含まれる割合を高くする効果を多少期待できる (表 7、9 参照)。
- 本稿で用いた、高頻度の語が話題であるという方法では、必ずしも良い結果が得られたとはいえない。その理由としては、頻度のばらつきのない文章の先頭部分とかなりばらつきの出てくる最後の部分と同一の方法で行なっていることが考えられる (表 7、9 参照)。
- 省略語の先行詞 (話題) が、以前に主格として出現した名詞句である割合は、日経サイエンスの記事で 75 % と高く、省略語補完の際かなり有効である。また、天声人語においては、その割合が 45 % であり、日経サイエンスの記事よりは小さいが、省略補完の際、省略語の先行詞の候補の中に正しい先行詞が含まれる割合を高くする効果を期待できる (表 8、10 参照)。
- 省略語の先行詞 (話題) が、以前に主格として出現した名詞句である割合は、天声人語より日経サイエンスの記事の方が高い。この結果も先の調査結果 [1] の裏付けになっている (表 7～10 参照)。

6 むすび

本研究では、先に提唱した省略に関する4つの談話的制約を用いて計算機で省略語を補完するという実験を行ない、それらの制約の有効性を検証した。それに加え、省略され得る情報と定義した話題を推定する2つの方法を提案し、実際の文章を用いて行なった調査の結果からそれらの方法が省略補完の精度の向上に有効であることを述べた。

今後の課題として以下の3つの事柄が考えられる。

1. 制約の使用法の改善

本稿では、制約に反する語を完全に省略語の先行詞の候補から除去するという制約の強い適用法で省略補完を行なったが、必ずしも良い結果が得られたとはいえない。そこで、制約によっては、1、2度制約に反しても省略語の先行詞の候補に残すという制約の緩やかな適用法などを採用するなどして、制約の使用法を改善する必要がある。

2. 話題を組み込んだ省略補完システムの構築

省略補完の際、話題を導入することで省略語の先行詞の候補中に正しい先行詞を残す割合を高くできるということが、明らかになった。しかしながら、現在の省略補完システムでは、話題を考慮していないので、話題を組み入れた省略補完システムを構築するとともに、そのシステムを実際に計算機で実験する必要がある。

3. 新しい制約の考案

話題も考慮に入れた省略補完システムでも、満足できる結果が得られなかった場合、さらに、省略に関する談話的制約を考案する必要がある。

謝辞

「角川類語新辞典」[4]を計算機可読の形で提供して頂き、その使用許可を頂いた

(株)角川書店に深謝する。

参考文献

- [1] 藤澤、増山、内藤：日本語文章における照応・省略現象の基本的検討、情報処理学会自然言語研究会研究報告、NL-86-6(1991)
- [2] 久野 章：談話の文法、大修館 (1978)
- [3] 情報処理振興事業協会技術センター：計算機用日本語基本動詞辞書 IPAL 解説編、辞書編 (Basic Verbs, 1987)、(Basic Adjectives, 1990)
- [4] 大野、浜西：類語新辞典、角川書店 (1981)
- [5] J.Hinds, S.Maynard and S.Iwasaki : Thematization staging assumed familiarity, and syntactic binding in Japanese, Perspectives on Topicalization : The case of Japanese wa. Amsterdam : J. Benjamins Publishers. (1987)
- [6] J.Hinds, S.Maynard and S.Iwasaki : Identifiability, scope setting, and the particle wa : A study from Japanese expository discourse, Perspectives on Topicalization : The case of Japanese wa. Amsterdam : J. Benjamins Publishers. (1987)
- [7] J.Hinds, S.Maynard and S.Iwasaki : Thematization as a staging device in Japanese narrative, Perspectives on Topicalization : The case of Japanese wa. Amsterdam : J. Benjamins Publishers. (1987)
- [8] 井上和子：旧い情報、新しい情報、月刊言語、Vol.8, No.10, pp.22-34(1979)
- [9] 井上和子：文-文法から談話文法へ、月刊言語、Vol.12, No.12, pp.38-46(1983)