

"名詞らしさ"と品詞の概念

塚脇 幸代

s-tsuka[at]dream.ocn.ne.jp

はじめに

名詞、動詞、形容詞などは、各々の形態に応じて与えられた名前であるが、実際の文においては、往々にしてその形態を逸脱する。たとえばサ変動詞語幹、形容動詞語幹が名詞と同等の役割を果たすことは、経験的によく知られた事象である。また、副詞的名詞、名詞的副詞などと呼ばれる範疇もある。本稿では、われわれが名詞と呼んでいるものや、名詞と同じ役割を担うと考えている語が、どの程度「名詞らしい」のかを測るために、機能語、とりわけ助詞類との関係をコーパスから獲得し、分析を試みる。

1. 方法

“名詞”的なものを収集するために、形態素解析器 MeCab 0.98 と辞書 UniDic 1.3.12 を用いる。解析対象のコーパスは、「日英新聞記事対応付けデータ(JENAAD)」(Utiyama and Isahara, 2003)の一対一文対応ファイルの日本語部分とする。¹

抽出条件は以下のとおりである。

(1) MeCab+UniDic の解析結果から、以下の品詞情報を持つものを、名詞連続をなすものとして取り上げる。これを要素 A とする。なお、()内は細分類または活用形である。UniDic では、サ変動詞語幹またはサ変名詞と呼ばれるものは名詞に、形容動詞語幹と呼ばれるものは形状詞に分類されている。

名詞
形容詞(語幹一般)
形状詞
代名詞

動詞(連用形一般)

連体詞

副詞

記号(文字)

接尾辞(名詞的/形状詞的)

接頭辞

(2) 名詞連続に後接する要素として、以下の機能語を取り上げる。これを要素 B とする。

助詞(格助詞)

助詞(係助詞)

助動詞「な」

助動詞「だ」

(3) 要素 A に後接する要素 B があれば、要素 A と要素 B を組み合わせ、[出現数]#[要素 A]@[要素 B]の形で抽出する。

抽出形式は要素 A を品詞とするものと、出現形とするものの二通りとする。

抽出形式1の例: 126561#名詞@を

抽出形式2の例: 6#打破@を

抽出用に用意したテキストの大きさは、149,528 行、形態素数 5,128,562 (文区切りを含む)である。

抽出は perl 5.10 を用いて、UTF-8 コードを扱う環境で行った。

2. 結果

抽出形式1による出力を基に、出現数降順でソートした結果を表1に、抽出形式2による出力結果の一部を表2に示す。要素 B に当たる機能語は 29 種類抽出され、全部で 115 のパターンが抽出された。要素 A が名詞のパターンが上位に来るのは直観に基づく予想に違わない。

それに対し、表中で下位にあるパターンは、解析ミスなども含まれていると予想されるが、「連体詞 @が」のパターンにカウントされた出現形を見ると、「同じ@が」であったり、興味深い例も観察される。

¹ 本文中の例文(斜体で示す)はこのコーパスからの引用である。

出現数	パターン	出現数	パターン	出現数	パターン	出現数	パターン
176490	名詞@の	811	副詞@と	34	動詞連用@は	2	形容詞語幹@の
126561	名詞@を	740	代名詞@の	31	代名詞@だ	2	名詞@～
112829	名詞@に	682	形状詞@と	23	名詞@ノ	2	名詞@モ
81800	名詞@が	471	副詞@に	21	形状詞@を	2	名詞@っか
71355	名詞@は	439	代名詞@から	18	接尾辞@こそ	2	動詞連用@こそ
52693	名詞@で	427	代名詞@と	18	形状詞@で	2	副詞@より
41237	接尾辞@の	306	名詞@こそ	14	動詞連用@も	1	連体詞@が
38809	名詞@と	256	代名詞@より	13	動詞連用@が	1	接頭辞@を
24733	接尾辞@に	213	副詞@だ	12	副詞@な	1	接頭辞@の
19163	名詞@も	192	副詞@は	11	記号@へ	1	接頭辞@と
15451	接尾辞@を	172	動詞連用@に	10	記号@で	1	形状詞@こそ
13765	接尾辞@が	159	接尾辞@より	10	名詞@ン	1	形状詞@から
13692	接尾辞@は	148	副詞@も	8	名詞@之	1	形容詞語幹@だ
13486	名詞@だ	123	記号@の	8	副詞@が	1	名詞@乃
13331	名詞@から	121	記号@は	7	名詞@ト	1	名詞@ニ
11408	形状詞@な	74	副詞@で	6	形状詞@が	1	名詞@ッ
7631	接尾辞@な	57	動詞連用@の	6	名詞@へ	1	名詞@にて
7317	接尾辞@と	54	代名詞@へ	5	記号@から	1	名詞@っと
6952	接尾辞@で	50	動詞連用@を	5	名詞@ヲ	1	動詞連用@へ
6884	名詞@へ	48	記号@が	4	記号@も	1	動詞連用@より
5177	名詞@な	46	記号@と	4	形状詞@も	1	動詞連用@な
4395	接尾辞@から	44	副詞@から	4	形状詞@は	1	副詞@ト
3653	接尾辞@も	43	記号@を	3	動詞連用@より	1	代名詞@じ
2501	代名詞@に	43	代名詞@な	3	記号@だ	1	や
2291	形状詞@だ	40	動詞連用@と	3	名詞@じゃ		
2208	代名詞@は	40	代名詞@こそ	2	動詞連用@だ		
1882	形状詞@の	39	記号@に	2	副詞@を		
1763	代名詞@を			2	記号@より		
1394	代名詞@が			2	接頭辞@も		
1181	名詞@より			2	接尾辞@之		
1180	代名詞@も						
1151	接尾辞@だ						
1071	代名詞@で						
1017	接尾辞@へ						
918	副詞@の						

表1: 抽出形式1による出力結果

パターン	出現形1	出現形2	出現形3	出現形4
1#名詞@っと	1#バァー@っと			
41237#接尾辞@の	1545#年@の	1191#後@の	1221#つ@の	486#力@の
918#副詞@の	105#かつて@の	24#突然@の	198#初めて@の	282#一層@の
1#接頭辞@の	1#双@の			
57#動詞連用@の	1#訴@の	1#育て@の	3#し@の	2#洗い直し@の
123#記号@の	8#K@の	9#P@の	14#A@の	1#v@の
13331#名詞@から	159#立場@から	12#決算@から	1#いつごろ@から	262#観点@から
740#代名詞@の	15#あなた@の	21#どちら@の	126#我々@の	22#いつ@の
1882#形状詞@の	206#当然@の	17#骨太@の	51#過度@の	1#あいにく@の
2#形容詞語幹@の	2#カビ臭@の			
1#連体詞@が	1#同じ@が			
1017#接尾辞@へ	113#化@へ	44#人@へ	2#権@へ	3#船@へ
6884#名詞@へ	21#開発@へ	5#フランス@へ	11#関係@へ	3#ミャンマー@へ
54#代名詞@へ	20#どこ@へ	8#そこ@へ	16#ここ@へ	8#それ@へ
1#動詞連用@へ	1#掘り@へ			
3#副詞@を	1#ぎくしゃく@を	1#つい@を	1#ことごとく@を	
7317#接尾辞@と	6#済み@と	5#強@と	263#覚@と	522#者@と
811#副詞@と	47#しっかり@と	1#ぎゅうぎゅう@と	1#とぅとぅ@と	2#だらだら@と

表2: 抽出形式2による出力結果(一部)

3. 考察

抽出形式2による出力を基に、要素 B に「が」「だ」「と」「な」「に」「の」「は」「も」「を」をもつ抽出パターンを用いて、多次元尺度構成法により得た機能語の分布を図1に示す。「が」と「を」が近い位置にプロットされている。両者が名詞のマーカーとして目されるべき裏づけとなるのであろうか。また「な」は他と特徴的な位置を占めることが推察される。

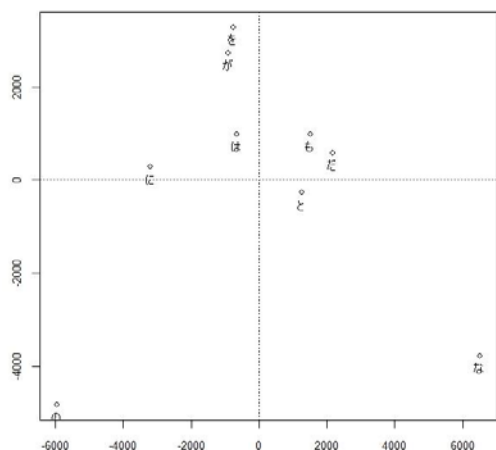


図1:多次元尺度構成法による機能語の分布

また、抽出形式1の出力結果に基づいた品詞の分布を図2に示す。座標上で名詞が突出している理由については、やはり名詞は名詞という結論も考えられるが、もう少し詳細を見るべきかも知れない。

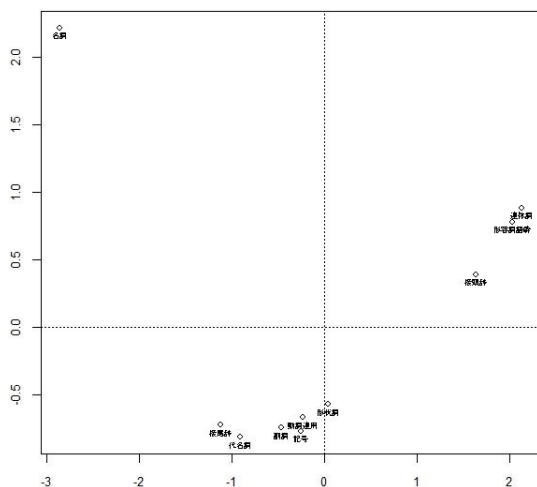


図2:多次元尺度構成法による品詞の分布

今回の出力結果を考察するにあたり心得ておかなければならないことは、この出力は(1)特定のコーパスから、(2)現在利用可能な形態素解析器の一つを用い、(3)その形態素解析器で利用可能な辞書の一つを用い、その情報に基づいて、獲得されているという点である。すなわち、コーパスが異なれば、得られる語彙も品詞も異なり、解析器や辞書が異なれば、得られる形態素の単位にも違いが生じ、解析の結果与えられる品詞の種類も異なるはずである。ただ、今回の出力をもって観察された、いくつかのパターンから、品詞の定め方についての示唆を得ることができる。前節で指摘した「連体詞@が」のパターンについて見ておく。

「連体詞@が」のパターンの出現頻度は1であり、その出現形は、「同じ@が」であった。元の文は以下のようなものであった。

これに対して、前年より上昇したのは、一・八%アップの島根をはじめ佐賀、山口の計三県だけで、前年と同じが岩手など東北地方を中心に六県だった。

また、「同じ」に後接する要素Bの機能語と頻度は表3のごとくであり、「が」が後接するパターン以外は、いずれも形状詞と解析されている。

	が	だ	と	の	計
同じ	1	37	7	1	46

表3:出現形「同じ」に後接する機能語

「形状詞@が」のパターンは6例ではあるが観察されるので、「(フレッシュ@が)」「好き@が」など解析器や辞書が接続を許していないというわけではないと見られる。「同じが」という文字列が、「形状詞」+「が」と解析されなかったのは、「同じ」が連体詞を持っており、形状詞より優先された結果であろうが、連体詞に格助詞「が」が後接するという現象は、文法理論上は考えられない。

「同じ」が連体詞となりえるとすれば、以下のように名詞が後接する場合であり、実際連体詞として解析される。

JR東海によると、同じホームに二本の新幹線を停車させた例はこれまでないという。

4. 議論

本稿の主旨に立ち戻ると、要素 A として抽出され

る品詞は、いずれも名詞句の要素となることを前提として選ばれている。連体詞や接頭辞は、名詞句の一部となりえても、直ちに助詞に後接されることは想定していない。その意味では、本手法により一つの予測しない結果を得たことになる。処理上はエラーとも見えるが、品詞のあり方を見直すきっかけともなりえる。このことを別においても、そもそも名詞として扱うべき品詞が多種にわたる状態は、文法理論の不完全さを知らしめるものであり、言語処理に負荷をかける一因ではないか。辞書が整備され、形態素解析の精度が高くなっていくにつれ、このような負荷は軽減されたかのように錯覚をおこさせるが、そこで使われている品詞とその体系には、大きな改良は行われてこなかったように思う。各方面への言語処理の応用が進む現状においては、処理を根底から覆しかねない品詞体系の再構築は冒険であるが、早晚必要になるであろう。

5. 関連研究

既存の品詞体系に問題を提起する研究 2 件について、言及しておく。

ひとつは水谷 (1994, 2001) の”語類立て”である。人手で収集した用例を元に、名詞とその周辺の品詞 (形容動詞や副詞) にいたるまで、機能語との接続関係や陳述の有無について調査され、新しい品詞体系を構築する試みが行われている。これより後に発表された参考文献 [9] と着想は共通する部分があるが、はるかに高品質でかつ綿密である。

もうひとつは、言語学の分野で取り上げられている”分散形態論”である。言語学の内部でも諸説がある理論であるが、これについては、田川 (2008) の分析を挙げておく。

言語学の分野で湧き上がった理論が認知され、実装されるに至るまでには多少の時間が必要になるかもしれないが、今後これらの研究が言語処理に良い影響を与えることを期待する。

おわりに

名詞らしさとはなにか、を明らかにするため、出力を行った。今後、得られた結果をいま少し詳細な

目で観察し、手法についても、対象とした品詞についても、さらに工夫をして分析を行ない、品詞体系の改善につながる現象を獲得することを目指す。

参考文献

- [1] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. URL <http://mecab.sourceforge.net/>.
- [2] 伝康晴, 小木曾智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵 (2007). 「コーパス日本語学のための言語資源: 形態素解析用電子化辞書の開発とその応用」, 『日本語科学』22 号 pp.101-122
- [3] Masao Utiyama and Hitoshi Isahara. (2003). Reliable Measures for Aligning Japanese-English News Articles and Sentences. ACL-2003, pp. 72--79.
- [4] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- [5] 中村純作 (2007). 「R を使った多変量解析入門」. 立命館大学国際文化研究所/アジアの英語教科書コーパス ワークショップ資料.
- [6] 水谷静夫, 星野和子 (1994). 「名詞から副詞までー語類の新しい枠づけ」. 『計量国語学』第 19 巻 7 号, pp.331-340.
- [7] 水谷静夫 (2001). 「束論から見た語類立て」. 『計量国語学』第 23 巻 3 号, pp.331-340.
- [8] 田川拓海 (2008). 「分散形態論による活用への統語論的アプローチー現代日本語における動詞連用形の形態統語論的分析」. 『筑波応用言語学研究』15 号, pp.59-72.
- [9] 塚脇幸代 (2007). 「対訳コーパスにおける品詞タグー名詞属性を持つ日本語品詞ー」. 言語処理学会第 13 回年次大会. PD1-2.