単語辞書を用いた英語品詞間の転換に関する調査

村田 真樹　進藤 三佳　馬 青　井佐原 均
独立行政法人 通信総合研究所
{murata,mshindo,qma,isahara}@crl.go.jp

1 はじめに

英語の単語には品詞の転換という現象がある [1]. これは接辞を付加せずに単語の品詞を変えることである．例えば，bottle という語は「瓶」という名詞であったが，「瓶詰にする」という動詞としても用いられるようになっている．本稿はこのような品詞転換の現象を，英和辞典などの単語辞書を用いて調査する．具体的には各単語が持つ品詞の種類を電子化された単語辞書から自動取得し，品詞の種類ごとにそれら品詞を複数持つ単語の個数などを調査して，どういう品詞とどういう品詞の間で品詞の転換が生じやすいかを調べる．その際，自己組織化マップを利用した調査結果の可視化も行なう．これらの調査結果は転換現象などの言語の歴史的変遷を調べる言語学 [2] の基礎データとして役に立つ可能性がある．

2 単語辞書を用いた数量的調査

本調査では 単語辞書としてジーニアス英和辞典 [3] のデータを利用した．この辞書の見出し語の総数は，74848 個であった．次に各見出し語 (単語) に出現する品詞のマークの数を計測した．その結果を表1に示す．この個数の算出の際は，例えば，名詞と形容詞の品詞を持つ単語の個数は，名詞の出現回数と形容詞の出現回数の両方に加算される．表からわかるように名詞の単語がもっとも多く次に形容詞が多い．また，この辞書では品詞のマークのつかない単語もあった．本研究の調査では品詞のマークのつかない単語は扱わず，少なくとも一つは品詞のマークのついている 56837 個の単語を対象とした．

次に二つの品詞を持つ単語の割合などを調べて，どういう品詞とどういう品詞の間で品詞の転換現象がおきやすいかを調べた．例えば，品詞 A と品詞 B の両方を持つ単語の個数が多ければ品詞 A と品詞 B の間で多くの転換現象が起きていることがわかる．

この調査結果を表2に示す．表の「二項以下のみ」は，品詞を三種類以上持つ単語を除いて調査した結果で，表の「すべて」は，品詞を三種類以上持つ単語を含めて品詞のマークをもつすべての単語で調査した結果である．「共起頻度」は，左の欄で示す二つの品詞をともに持つ単語の個数である．ただし，表の「すべて」の場合などで生じる，品詞を三種類以上持つ単語については，2/n の重み[1]をかけて用いる．n は，その単語が持つ品詞の個数である．これは例えば品詞を A,B,C の三種類持つ場合，転換現象は A→B→C, A→C→B, B→A→C, B→C→A, C→A→B, C→B→A の六つの順序で生じている可能性があるが，そのうち，A,B の品詞の組の間で転換が起きている割合が 2/3(=4/6) であるように，三種類以上の場合は対象とする二つの品詞の間で転換が起きている確率が 1 よりも下がるためこのような計算を行なう．表の「共有率」は例えば品詞 A と品詞 B の間でのものならば以下の式で与えられる．

$$\text{共有率} = \frac{F_{a,b}}{F_a + F_b - F_{a,b}} \quad (1)$$

表 1: 品詞の出現回数

品詞	出現回数	出現率
名詞	37158	0.4964
形容詞	14279	0.1908
動詞	8766	0.1171
副詞	5301	0.0708
間投詞	302	0.0040
連結形	242	0.0032
接尾辞	215	0.0029
接頭辞	157	0.0021
前置詞	145	0.0019
代名詞	126	0.0017
接続詞	71	0.0009
助動詞	32	0.0004
品詞付単語総数	56837	0.7594
単語 (見出し語) 総数	74848	1.0000

[1] 一般には，m 個の品詞の共起頻度を調査するときの n 個の品詞を持つ単語の重みは $(n+1-m)/{}_nC_m$ となる．${}_nC_m$ は n 個のものから m 個のものを選ぶ場合の数．

表 2: 品詞二項間の共有性の調査結果

品詞二項の組	二項以下のみ			すべて		
	共起頻度	共有率 (%)	予測比	共起頻度	共有率 (%)	予測比
名詞 形容詞	**3061**	**6.53**	0.35	**3494.9**	**7.29**	0.37
名詞 動詞	**4177**	**10.33**	0.78	**4514.6**	**10.90**	0.79
名詞 副詞	46	0.11	0.01	**255.6**	0.61	0.07
名詞 間投詞	**57**	0.16	0.39	102.8	0.28	0.52
名詞 前置詞	5	0.01	0.09	24.9	0.07	0.26
名詞 代名詞	9	0.02	0.13	14.8	0.04	0.18
名詞 接続詞	2	0.01	0.07	6.8	0.02	0.15
名詞 助動詞	2	0.01	0.11	4.7	0.01	0.22
形容詞 動詞	**276**	**1.28**	0.14	**570.1**	**2.54**	0.26
形容詞 副詞	**412**	**2.28**	0.35	**619.7**	**3.27**	0.47
形容詞 間投詞	6	0.04	0.11	18.3	0.13	0.24
形容詞 前置詞	8	0.06	0.37	27.6	0.19	0.76
形容詞 代名詞	13	0.10	0.52	23.0	0.16	0.73
形容詞 接続詞	1	0.01	0.09	5.8	0.04	0.33
動詞 副詞	1	0.01	0.00	70.4	0.50	0.09
動詞 間投詞	12	0.14	0.37	51.3	0.57	1.10
動詞 前置詞	2	0.02	0.15	6.9	0.08	0.31
動詞 代名詞	1	0.01	0.07	2.0	0.02	0.10
動詞 接続詞	0	0.00	0.00	1.8	0.02	0.17
動詞 助動詞	6	0.07	1.46	8.7	0.10	1.76
副詞 間投詞	4	0.08	0.20	14.7	0.26	0.52
副詞 接尾辞	2	0.04	0.11	2.0	0.04	0.10
副詞 前置詞	22	0.44	**2.82**	47.7	0.88	**3.53**
副詞 代名詞	4	0.08	0.44	15.2	0.28	1.29
副詞 接続詞	17	0.34	**4.13**	28.3	0.53	**4.28**
間投詞 代名詞	1	0.31	**2.44**	1.5	0.35	2.24
間投詞 接続詞	0	0.00	0.00	0.7	0.18	1.77
連結形 接頭辞	1	0.25	1.47	1.0	0.25	1.50
接尾辞 接頭辞	1	0.27	1.66	1.0	0.27	1.68
前置詞 代名詞	0	0.00	0.00	1.3	0.49	**4.15**
前置詞 接続詞	7	**5.43**	**93.72**	15.5	**7.73**	**85.57**
代名詞 接続詞	1	0.67	**11.57**	3.3	1.72	**21.18**
調査単語総数	56004			56837		

ただし，$F_{a,b}$ は品詞 A,B の共起頻度，F_x は品詞 X の頻度 (出現回数) である．すなわち，品詞 A,B いずれかを持つ単語に対する品詞 A,B の両方を持つ単語の割合を意味する．表の「予測比」は例えば品詞 A と品詞 B の間でのものならば以下の式で与えられる．

$$\text{予測比} = Er(A,B) = \frac{F_{a,b}}{(F_a/Total)(F_b/Total) * Total} \quad (2)$$

ただし，$Er(A,B)$ は後の式で用いる品詞 A,B の予測比を意味する記号で，Total は調査対象の単語の総数

表 3: 品詞三項間の共有性の調査結果

品詞三項の組	三項以下のみ		
	共起頻度	共有率 (%)	予測比
名詞 形容詞 動詞	**362**	**0.697**	0.26
名詞 形容詞 副詞	**198**	**0.376**	0.23
名詞 形容詞 間投詞	4	0.008	0.09
名詞 形容詞 前置詞	6	0.012	0.31
名詞 形容詞 代名詞	3	0.006	0.16
名詞 形容詞 接続詞	1	0.002	0.10
名詞 動詞 副詞	**27**	0.059	0.05
名詞 動詞 間投詞	**49**	**0.119**	1.74
名詞 動詞 前置詞	1	0.002	0.08
名詞 動詞 代名詞	1	0.002	0.08
名詞 動詞 助動詞	4	0.010	1.25
名詞 副詞 間投詞	3	0.007	0.18
名詞 副詞 前置詞	7	0.017	0.98
名詞 副詞 代名詞	2	0.005	0.28
形容詞 動詞 副詞	**17**	**0.064**	0.09
形容詞 動詞 間投詞	1	0.004	0.09
形容詞 副詞 前置詞	8	0.043	**2.94**
形容詞 副詞 代名詞	9	0.048	**3.34**
形容詞 副詞 接続詞	2	0.011	1.48
動詞 前置詞 接続詞	1	0.011	**52.80**
副詞 間投詞 接続詞	1	0.018	**37.40**
副詞 前置詞 接続詞	7	**0.132**	**618.22**
調査単語総数	56718		

である．これは，品詞 A,B が独立に出現すると仮定した場合品詞 A,B の頻度のみから予測した品詞 A,B の共起頻度に対する実際の品詞 A,B の共起頻度の値の比である．つまり，この値が大きいほど予想される個数に比べて多くの共起が生じていることになる．ところでこの式は品詞二項の組を対象とする場合は log_2 をつけると自己相互情報量と等しいものとなる．

また，同様の調査を品詞三項の組でも行なった．その結果を表3に示す．紙面の都合で，品詞三項の組での実験では「すべて」の場合の結果は示さず，品詞を四種類以上持つ単語を省いた調査である「三項以下のみ」の結果しか示さない．「共起頻度」は三つの品詞がともに出現した単語の個数で「共有率」は以下の式で

$$\text{共有率} = \frac{F_{a,b,c}}{F_a + F_b + F_c - F_{a,b} - F_{b,c} - F_{c,a} + F_{a,b,c}} \quad (3)$$

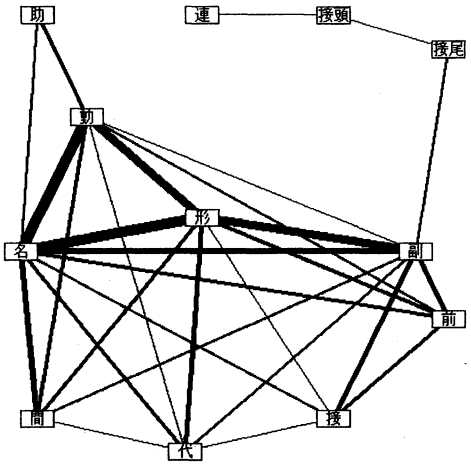

図 1: 共起頻度に基づく品詞マップ

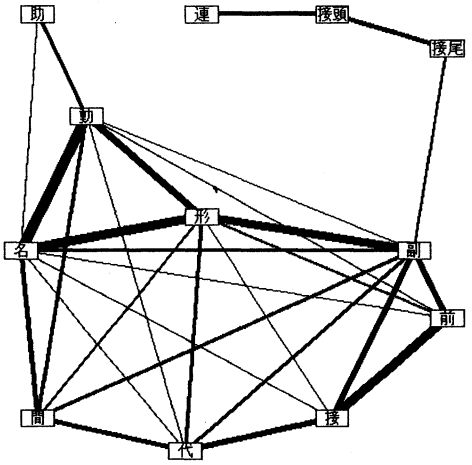

図 2: 共有率に基づく品詞マップ

「予測比」は以下の式で与えられる．

$$\text{予測比} = \frac{F_{a,b,c}}{(F_a/Total)(F_b/Total)(F_c/Total) * Total} \quad (4)$$

二つの調査結果の表2，表3では，それぞれの尺度で上位5個の品詞の組の数値を太字で示している．また表にない品詞の組は共起頻度が0であったものである．

本稿では品詞間の共有性の調査には「共起頻度」「共有率」「予測比」しか用いていないが，その他の尺度も共起頻度や個々の品詞の個数，調査単語総数から求まるものならば，本稿と同様，容易に単語辞書から算出可能である．また，本稿ではどういう尺度が調査に役に立つかがわからなかったので，共有性を調べるときの基本的な尺度の「共起頻度」「共有率」を用い，また「共起頻度」だと頻度の大きい品詞の値が大きくなる欠点があり，また「共有率」だと比較する品詞間で頻度に差がある場合値が小さくなる欠点があるため，品詞間で頻度に差があっても影響されない「予測比」も用いた．

3 調査結果の可視化

次に先の数量的な調査結果を二次元のマップ上に表現した．これは先の数量的な調査結果を可視化することで，より効果的な調査結果の提示を試みるものである．この調査は，表2の「二項以下のみ」のデータで行なった．その結果を，図1〜図3に示す．図での線の太

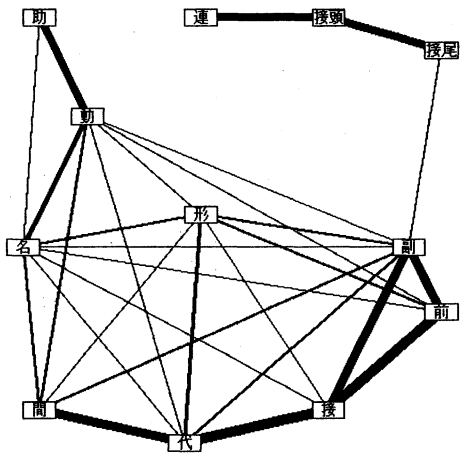

図 3: 予測比に基づく品詞マップ

さはその結ぶ品詞間での共起頻度，共有率，予測比の値の大きさに対応しており大きいものほど太い線で結ばれるようになっている．各品詞の表記は接頭辞，接尾辞のみ最初の二文字を使い，他は最初の一文字のみを用いている．各品詞の配置には Kohonen の自己組織化マップ (SOM_PAK) を用いた [4]．またデータとしては予測比のものを用いた．これは共起頻度，共有率を用いてできたマップよりも見やすそうだったからである．SOM の入力の行列は文献 [5] を参考にして各品詞間の非類似度 (距離) を意味する行列を与えた．行列の非対角要素の品詞 A と品詞 B の非類似度としては以

下の値を用い，

$$\frac{1}{2} - \frac{log_2 Er(A,B)}{2(1+|log_2 Er(A,B)|)} \quad (5)$$

行列の対角要素は0とした．ただし，$Er(A,B)$は式2で与えられる品詞A,Bの予測比である．この式にはそれほど理論的根拠はなく予測比が0のとき1，1のとき1/2，∞のとき0となる式として選んだ．13×13の2次元配列のSOMを用い，整列フェーズでは学習回数を10,000に学習率の初期値を0.1に近傍半径を13に，微調整フェーズでは学習回数を1,000,000に学習率の初期値を0.01に近傍半径を7にして自己組織化を行なった．その結果で間投詞と代名詞の位置を交換し各品詞を上下左右にすこし人手で動かしたものを品詞の配置として利用した．この品詞の配置の移動は品詞間に線を結ぶ際三つ以上のものが一直線にならぶと各線がどの品詞間のものかがわからないためで，そうならないように移動した．

4 考察

表2より名詞 - 動詞の組で共起頻度が大きくこの品詞の組で多くの転換現象がおこっていることがわかる．また，名詞 - 形容詞の組でも共起頻度より転換現象が多いことがわかる．文献[1]には，従来あまり生産的でないとされていた名詞 - 形容詞の転換現象はそれほど少なくなくむしろ多いと，記述されている部分があるが，そういう名詞 - 形容詞の転換現象の多いことが本稿の調査でもすぐにわかるのである．また本稿では示さないが，本稿の単語辞書を用いた調査方法では実際にその転換をしている単語の具体例も共起頻度の個数分容易に取り出すことができる．ただし，本稿の調査では転換の方向性はわからない．例えば，ある単語が名詞と形容詞の二つの品詞をもっていても，どちらからどちらの品詞へ転換したかはわからない．これには注意しておいてほしい．

共有率，予測比を見ていくとまた違った品詞の組で強い関係があることがわかる．予測比が大きいものとしては，副詞 - 接続詞や前置詞 - 接続詞がある．これらは品詞間の機能的な類似性が影響しているのであろう．

また，表2では「すべて」の場合も示しているが，「すべて」の場合は共起頻度に確率的な確からしさで重み付きで頻度を加えているものがあるので，厳密な値でなく，推測した値である．この結果を使う場合は注意が必要である．とはいえ，逆に「二項以下のみ」の場合はすべての単語を用いていないという問題がある．

表3より三項の組の強さとしては，共起頻度では名詞 - 形容詞 - 動詞が強く，予測比としては副詞 - 前置詞 - 接続詞が強いとわかる．ところで，これは可視化した図1，図3を見てもわかることで，それぞれの図で太字の三角形を見ると上の組であることがわかる．可視化した図は品詞三項のつながりの強さなど，より広い品詞間のつながりを考察するのに役立つのである．また，図での配置も興味深く，助動詞と動詞が近くに，また，連結形，接頭辞，接尾辞が近くに配置されるなどの構造が自動で得られている．

可視化した図の作成には自己組織化で作ったマップを基礎として用いたが，この品詞の配置も人手で行なうのは困難である．自己組織化マップの考え方は本稿のような調査研究の可視化にも役に立つのである．また，本稿では品詞をマッピングするだけでなく各品詞間を線でつなげ各品詞間のつながりの強さを線の太さで表現したが，これにより品詞間の関係もより明確に示すことができた．

5 おわりに

本稿では単語辞書を用いて英語品詞の転換現象の効率的な調査を行なった．本稿では詳細な調査結果の表とそれを可視化した図を示すことができた．これらの結果は転換現象などの言語の歴史的変遷を調べる研究の基礎的なデータとして役に立つものと思われる．特に可視化した図は，じっくりと見ているうちになにか深いものが見えてきそうである．本稿では調査手続きと調査結果の詳述につとめた．この調査結果に対する言語学的な考察は今後の課題である．

参考文献

[1] 竝木崇康，語形成，新英文法選書，第2巻，(大修館書店，1985).

[2] Mika Shindo, A panchronic approach to semantic extensions of visual expressions in English, *Papers from the twentieth National Conference of The English Linguistic Society of Japan (JELS 20)*, (2003).

[3] 小西友七（編），ジーニアス英和辞典第2版，(大修館書店，1996).

[4] Teuvo Kohonen, *Self-organizing maps, 2nd Edition*, (Springer, 1997).

[5] 馬青，神崎享子，村田真樹，内元清貴，井佐原均，日本語名詞の意味マップの自己組織化，情報処理学会論文誌，Vol. 42, No. 10, (2001), pp. 2379–2391.