テキストにおける多義語の意味の集中度

山崎　誠（国立国語研究所言語資源研究系）†

The Concentration Ratio of Polysemous Senses in Japanese Text

Makoto Yamazaki (Dept. Corpus Studies, NINJAL)

要旨

テキストにおいて多義語の意味が特定のひとつに集中して用いられる傾向があることは既に Gale et al.(1992)らによって指摘されているが、先行研究では特定の何語かを取り上げて、その出現傾向をさぐるものがほとんどであった。本研究はテキストに現れた全ての名詞について、意味の集中の度合いを測定しその結果を報告するものである。その結果、当該テキストの話題によって特定の意味に集中する傾向はあるものの、実質的な意味と形式的な意味が共存している場合は、複数の意味が生じやすいことが分かった。このことは、テキストにおける結束性は主に語彙的な部分で働き、文法的には働かないことを意味していると解釈される。

１．語彙的結束性

語彙的結束性（lexical cohesion）は、テキストを成立させる重要な条件として Halliday & Hassan(1976)によって提唱され、テキストにおける同一の語の繰り返し使用などについて計量的研究が行われてきた。例えば、多義語については、山崎(2010:30)において「テキストにおける多義語の意味実現が一つの意味に偏りやす」く、その偏りは「出現間隔が近いほど起こりやすい」ことが指摘されている。多義語の意味実現は、語彙的結束性という概念を用いず、自然言語処理の観点からの研究が早くから行われており、上掲の Gale et al.(1992)に対して、細かな意味の違いを考慮すれば、複数の意味実現が見られるという指摘もある（Krovetz(1998)）。本稿は従来のアプローチとは違い、テキスト全体における多義語の意味分布を探る試みである。そのことにより、語彙的結束性のあり方を把握すると同時に、多義の意味分布から見たテキストの特性についても考察する手がかりとする。

２．データ

本稿では、『現代日本語書き言葉均衡コーパス』（以下、BCCWJ と略す）の中の図書館書籍（LB）のデータを利用した。「BCCWJ 短単位語数」のページ[1]で公開されているファイル BCCWJ_WC_SUW_v10.xlsx を利用し、LB で空白、記号、補助記号を除外した可変長部分の短単位数の中央値が 2360 であることから、2360±50 短単位の範囲から NDC の異なるサンプルをランダムに抽出した。使用したサンプルの概要を表 1 に示した。LBd0_00003 は、クロワッサンのレシピを説明したエッセイ、LBj4_00045 はアリの生態を解説した科学書、LBn9_00088 はチェスのプレイヤーを主人公にした翻訳小説である。

† yamazaki@ninjal.ac.jp

[1] https//maro.ninjal.ac.jp/wiki/index.php?BCCWJ%2F%E7%9F%AD%E5%8D%98%E4%BD%8D%E8%AA%9E%E6%95%B0

表 1 使用したサンプル

サンプル ID	NDC	タイトル	著者	出版社	出版年	短単位数
LBd0_00003	0 総記	空飛ぶフランスパン	金子郁容(著)	筑摩書房	1989	2403
LBj4_00045	3 自然科学	アリはなぜ一列に歩くか	山岡亮平(著)	大修館書店	1995	2310
LBn9_00088	9 文学	ディフェンス	ウラジーミル・ナボコフ(著),若島正(訳)	河出書房新社	1999	2373

３．方法

３．１　形態素解析について

形態素解析は BCCWJ に付与されている短単位の情報をそのまま用いた。エラーの修正は行っていない。

３．２　多義語の認定について

該当の語が多義かどうかについては『三省堂国語辞典第七版』（以下、『三国』と略す）によった。この辞書は比較的多義を認定している可能性が高いことから、より多くの分析対象が抽出できることを期待して採用したものである。

３．３　語義の数え方

各サンプルに出現した各短単位の中から頻度 2 以上の名詞（普通名詞）[2]を抜き出し、個々の使用例ごとに『三国』のどの語義に相当するかを判断した。多義のレベルは①②などの丸付き数字のレベルで判断し、それ以下の(a)(b)などの区分は区別しない。また、[一][二]などで示された品詞の区分が異なる場合は、異なる語義とみなした。短単位と『三国』とで語のまとめ方が異なる場合、短単位を基準にした。例えば、短単位では動詞「のる」は『三国』の「乗る」と「載る」を合わせたものに相当する。したがって、「乗る」の 15 個の語義と「載る」の 3 個の語義を合わせたものを短単位の「のる」に対応させた。

以下のような事例については分析の対象から除外した。

(1)誤解析と認められるもの[3]。

(2)該当する見出し語が『三国』にないもの[4]。

(3)該当する語義が『三国』にないもの[5]。

対象となった語数を表 2 にまとめた。対象とした多義普通名詞（表 2 のいちばん下の 2 行）をサンプル全体に対する比率で見ると、LBd0_00003 は、それぞれ 13.9%（延べ）と 13.5%（異なり）、LBj4_00045 では、12.1%（延べ）と 11.8%（異なり）、LBn9_00088 では、延べ・異なりともに 7.6%である[6]。

---

[2] 名詞のサブカテゴリのうち、固有名詞と数字は多義性がほとんどないと考えられることから除外した。

[3] 例えば、LBd0_00003 において「縁（えん）」と解析された 2 語は「ふち」ないしは「へり」となるものであったため、分析から除外した。

[4] 例えば、LBj4_00045 において「大蟻」「キャピラリー」「クロマトグラフ」「巣内」「侍蟻」「他種」「吐き戻し」「山蟻」などが『三国』の見出しにないため、分析から除外した。実際上、これはら多義の可能性が少ないので分析に与える影響は少ないと考えられる。

[5] 例えば、LBj4_00045 において、多義語の例として「手」4 例が出てくるが、これらは、慣用句「手に入れる」「手を煩わす」の一部であったり、複合語「お手の物」の一部であったりするため、該当する語義がなく、分析から除外した。

[6] ちなみにこの比率の差を多重比較したところ、延べ語数、異なり語数ともに LBd0_00003 と

表 2　対象とした多義語

語数	LBd0_00003	LBj4_00045	LBn9_00088
延べ語数	2403	2310	2373
異なり語数	594	524	683
度数 2 以上の普通名詞（延べ）	511	564	249
度数 2 以上の普通名詞（異なり）	125	113	76
度数 2 以上の多義の普通名詞（延べ）	362	341	249
度数 2 以上の多義の普通名詞（異なり）	91	75	76
対象とした多義普通名詞（延べ）	335	279	181
対象とした多義普通名詞（異なり）	80	62	52

３．４　語義の集中度

語義の集中度は、ある語についてサンプル中に出現した語義数（異なり[7]）をその語の持つ、可能性としての語義数（『三国』の語義数）で割った値を 1 から引いた値とした。すなわち、語義を 3 つ持つ語がサンプル中で 1 つの語義でしか使われなかった場合、1−(1/3)で、0.667 となる。もし、3 つの語義が全部使われていれば、1−(3/3)となり、集中度は 0 となる。語義が 1 つしかない語については集中度を算出しない。

４．結果

４．１　語義数の分布

図 1 は各サンプルにおける普通名詞における、可能性としての語義数の分布である。前述のように語義数は『三国』の語義数に拠っている。3 つとも似たような L 字形をしているが、LBn9_00088 はややゆるやかなカーブになっているのが特徴的である。

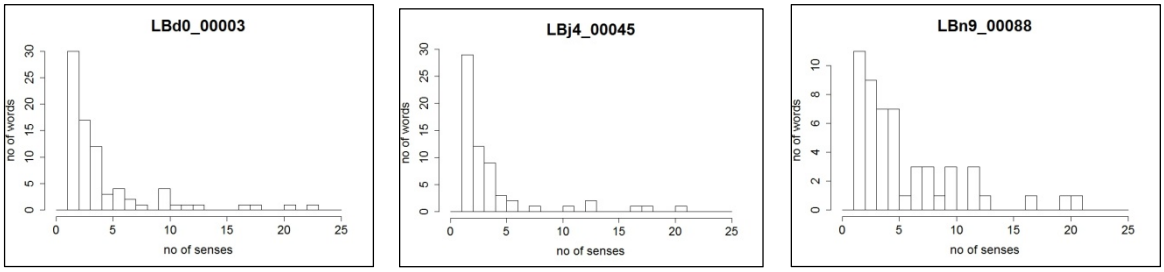

図 1　多義語の語義数の分布（『三国』の語義数の分布）

これに対して、実際に出現した語義数の分布が表 3 である。どのサンプルでも語義数 1 がいちばん多いが、LBn9_00088 は語義数 2 も割合が他より高くなっている。

表 3　出現した語義数の分布

語義数	LBd0_00003	LBj4_00045	LBn9_00088
1	69	52	38
2	10	7	12
3	1	2	-

LBj4_00045 との組み合わせは 5%水準で有意差がなかったが、そのほかの組み合わせは 5%水準で有意差が認められた。

[7] この指標は延べ語数を考慮していない。したがって、度数 10、語義数 2 の多義語があったとして、語義の分布が 9,1 の場合も 5,5 の場合は同じ値(0)になる。

4	-	1	1
5	-	-	1

４．２　集中度の分布

図 2 に各サンプルにおける多義の集中度の分布を示した。LBd0_00003 と LBj4_00045 はほぼ同じ形の分布を示しているが、LBn9_00088 は分布の形が異なっている。この差が何に由来するのかは不明であるが、LBn9_00088 が小説であることが関係している可能性を指摘しておく。

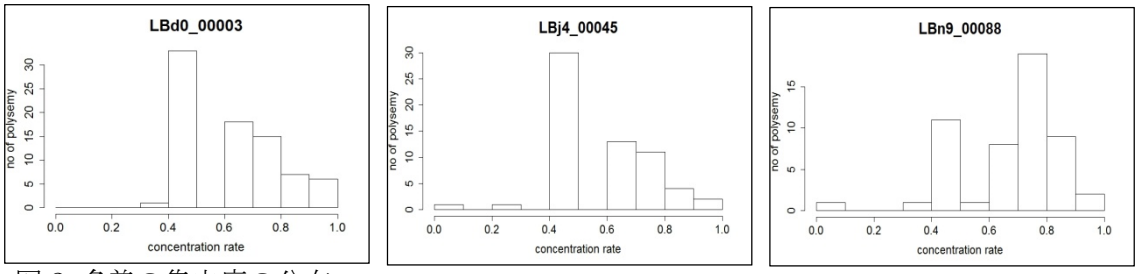

図 2 多義の集中度の分布

サンプル全体の集中度を表 4 に示した。平均値でみても、中央値でみても、LBj4_00045＜LBd0_00003＜LBn9_00088 の順になっている。事例数が少ないので憶測の域を出ないが、サンプル全体の集中度が当該テキストを特徴付ける指標になる可能性がある。

表 4　サンプル全体の集中度

代表値	LBd0_00003	LBj4_00045	LBn9_00088
平均値	0.649	0.606	0.693
中央値	0.667	0.500	0.750

表 5～7 は各サンプルにおける頻度 5 以上の語についての調査結果である。

表 5 語義の分布（LBd0_00003）頻度 5 以上

頻度	語彙素	『三国』語義数	出現語義数	内訳[8]	集中度
15	図	5	1	③	0.800
12	回	6	2	[二]①×5，[二]②×7	0.667
12	粉	2	1	①	0.500
12	方向	3	1	①	0.667
11	事	21	2	[一]①×4，[一]⑥×7	0.905
11	層	4	2	[一]①×7，二×4	0.500
9	板	5	1	①	0.800
8	ゲーム	4	1	②	0.750
8	センチ	2	1	②	0.500
8	作り	7	1	[二]①	0.857
8	度	10	2	[一]⑨×2，二×6	0.800
8	時	13	1	⑪	0.923
7	パン	2	1	①	0.500

[8] 内訳の欄の記号は『三国』の語義に対応する。×のあとの数字は出現数（token）である。ただし、出現語義数が 1 の場合、および、×1 の場合は省略した。

7	水	4	1	①	0.750
6	後	12	1	[一]②	0.917
6	固まり	2	1	②	0.500
6	時間	6	1	③	0.833
6	操作	2	1	①	0.500
6	東西	3	1	[一]①	0.667
6	中	8	2	①×3，④×3	0.750
5	上	10	2	[一]①，[一]②×4	0.800
5	三角	2	1	①	0.500
5	日	11	1	⑤	0.909
5	侭	4	1	③	0.750
5	ミリ	2	1	②	0.500

表 6 語義の分布（LBj4_00045）頻度 5 以上

頻度	語彙素	『三国』語義数	出現語義数	内訳	集中度
29	事	21	3	[一]6×24,[一]9,[一]⑪×4	0.857
17	奴隷	2	1	①	0.500
16	巣	3	1	①	0.667
13	種	3	1	①	0.667
11	炭化	2	1	①	0.500
10	相手	3	1	①	0.667
10	成分	2	1	①	0.500
10	物	17	4	[一]①×5,[一]⑥×2,[一]⑦,[一]⑬,除外 1	0.765
9	仲間	2	1	②	0.500
8	違い	3	1	一×7,除外 1	0.667
8	物質	2	1	②	0.500
8	分析	2	1	②	0.500
7	為	4	3	[一]①×5,[一]②,[二]	0.250
5	コロニー	4	1	③	0.750
5	自分	5	1	一	0.800

表 7 語義の分布（LBn9_00088）頻度 5 以上

頻度	語彙素	『三国』語義数	出現語義数	内訳	集中度
21	事	21	5	[一]①×8,[一]⑥×7,[一]⑨×4,[一]⑯,[一]⑰	0.762
19	娘	2	1	①	0.500
10	物	17	4	①×4,⑦×2,⑧×2,⑩,除外 1	0.765
8	時	13	2	⑪×6,⑬×2	0.846
6	声	5	1	①	0.800
6	二人	2	1	①	0.500
5	頭	12	2	①×2,⑤×2,除外 1	0.833
5	側	3	1	①	0.667

４．３　複数の語義で出現した語

各サンプルにおいて出現語義数が 2 以上の語は次の通りである。[] 内の数字は［可能性

としての語義数/出現語義数］である。

LBd0_00003：
回［6/2］、事（こと）［21/2］、層［4/2］、度（ど）［10/2］、中（なか）［8/2］、頃［4/2］、所（ところ）［18/2］、訳（わけ）［10/3］、最高［3/2］、地方［4/2］
LBj4_00045：
事（こと）［21/3］　物（もの）［17/4］　為（ため）［4/3］　所（ところ）［18/2］　情報［4/2］　筈（はず）［6/2］　以後［2/2］　程度［4/2］　時（とき）［13/2］　中（なか）［8/2］
LBn9_00088：
事（こと）［21/5］　物（もの）［17/4］　　時（とき）［13/2］　頭（あたま）［12/2］　後（あと）［12/2］　中（なか）［8/2］　癖（くせ）［4/2］　間（あいだ）［8/2］　顔（かお）［10/2］　最近［3/2］　姿［5/2］　全て［2/2］　展開［8/2］　訳（わけ）［10/2］

これらを見ると分かるように、複数の語義で出現した語のほとんどが抽象的な語であることが分かる。紙幅の関係で用例は省略するが、事（こと）、物（もの）、訳（わけ）、為（ため）、所（ところ）、筈（はず）など、意味が形式化して機能語に近い用法を持つ語が多い。また、「全て（名詞・副詞）」「以後（名詞、造語成分）」のように、品詞が違うことで多義として挙がっている語がある。一方、今回観察した 2400 短単位程度のテキストでは具体的な意味を持つ普通名詞で多義的に使用されている語はなかった。このことは、語彙的結束性は語彙的には強い制約としてテキストの成立条件となっているが、文法的にはその制約は弱いのではないかということが推測される。

５．まとめと今後の課題

本稿で観察したのは、普通名詞だけであったが、テキスト中では約 7 割～8 割の多義語が特定の意味でのみ使用されていることが確認された。その例外となっていたのは、ほとんどが文法的な意味での使用に関わるものであった。したがって、文法的な意味は語彙的結束性に関与する度合いが低いことを示唆した。今後はサンプル数を増やすとともに、動詞、形容詞などの他の品詞における多義の実現傾向、またテキストにおける使用頻度と出現語義数との関係も視野に入れて分析を行う予定である。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「コーパス日本語学の創成」による研究成果の一部である。データとして利用した BCCWJ は、国立国語研究所のプロジェクト及び文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18～22 年度、領域代表者：前川喜久雄）による補助を得て構築したものである。

参考文献

Gale,William A. et als.(1992), “One sense per discourse”, Proceedings of the workshop on Speech and Natural Language, pp.233-237, Harriman, NY.

Krovetz, Robert.(1998), “More than One Sense per Discourse”, Proceedings of the ACL-SIGLEX Workshop (Senseval)

Halliday, M.A.K. and Hasan, R.(1976) Cohesion in English. Longman.

山崎誠(2010),テキストにおける多義語の意味実現の傾向,計量国語学会第 54 回大会予稿集,pp25-30.