テキスト・コーパスからの語構成要素の語彙論的特徴の復元について

影浦 峡
国立情報学研究所 人間・社会情報研究系
kyo@nii.ac.jp

要約

語彙論は、(理論的仮説としてにせよ) 語彙空間の自律性を前提とする。一方、言語の一次的なデータは広義のテキストであり、辞書のような語彙論的データは二次的であるとの感を否めない。本研究では、一次的言語資料であるテキスト・コーパスから、テキスト空間における語や語基の特徴量ではなく、語彙空間における語基 (語構成要素) の特徴量を復元するための手法を検討する。考慮しなくてはならない点は、最低限、(1) 語彙空間域に応じて談話要因を除外するためのテキストのサンプリング、(2) 得られたテキスト・コーパスからの出現頻度要因の除外、(3) 複合語の結束構造があるが、本研究では、特に (2) に焦点をあて、語彙論的量的指標を導入する。それに基づき、分野依存コーパスを用いていくつかの語基の特徴量を比較検討する。

1 はじめに：語彙論の領域

ここでは語彙論を社会的に共有され (及び／或いは) 心理的に登録される語の集合としての語彙を対象とする研究と定義する[1] 。また、我々は、語彙論においてテキスト空間に対する語彙空間の実在性・自律性を仮定する。非歴史的語彙論の対象は、それゆえ、特定共時体における具体的な対象である (cf. 前田 1989)。とはいえ、「具体的な対象」たる語彙の研究が、現に存在する語の有限集合のみを博物学的な記述対象として満足するわけではない。語彙論の「具体性」は、わかりやすく示すならば、次のように規定できる。

(1) 語彙論は「許容可能な形式」にではなく**具体的な単位＝形式の現実的存在可能性**に関わる。
(2) 語彙の構造的予測は解釈上時間軸上の現実性と関連づけられる。

従って、例えば、「"a very long reddish brown extremely stinky hard nose"が許容される形式であるかどうか」は、語構成論が関知する問題であったとしても、語彙論が関知する問題ではない[2] 。

現実的存在可能性を対象とする限り、実際に発話／記録された言語データ[3] に依拠する必要がある。現実のデータは、一次的にはテキストであるから、そこから語彙論上の特徴を求める課題が語彙論に課されることになる。この枠組みでテキストから単位の語彙論上の量的特徴を求める場合、(1) 語彙空間域に応じて談話要因を除外するためのテキストのサンプリング、(2) 得られたテキスト・コーパスからの出現頻度要因の除外、(3) 複合語の結束構造、などを考慮しなくてはならない。本稿では、(2) を考える。

[1] 語彙論は英語の lexicology に相当するが、例えば Geeraerts (1994) や McCarthy (1991) などの lexicology に関する解説では、本研究の定義とは対照的に、語を巡る議論ではなく語彙を対象とするという観点も、そのための方法論を検討する視点もほとんどない。

[2] これが「存在可能っぽい語であるかどうか」は語彙論の問題であるが、語彙論の問題を考えるにあたり、現実の語彙には存在しないし存在したこともないこのような単位を持ち出すのは無意味で馬鹿げている。

[3] Foucault (1969) のモニュマン／ドキュマンも参照。

2 語構成要素の語彙空間での特徴量

本研究では、「語構成要素」(語基) の主要部 (head) としての語彙論上の特徴量 (造語力) を検討する。そのために、まず統計量を取る対象として統合的な単位 (ウィンドウ) を「語」(単純語・複合語) に限定する。つまり、テキストを語の延べリストと捉える (図1＜太字＞)。

＜**理論/物理**＞と＜**教育/心理**＞は、＜**観察**＞できる＜**範囲**＞では似通っており、＜**計算/機/科学**＞は＜**理論/物理**＞より高い＜**値**＞で＜**推移**＞している。

図1 テキスト・コーパスにおける着目点

図1で、「物理」、「心理」、「科学」、「観察」等の語彙論的「重み」を考えることが本研究の課題である。

最も基本的な指標は延べ頻度である：

$f(i,N)$:要素 w_i の延べ語数 N のコーパス中での頻度

これは、要素のディスコース上での使用度を表す。

また単純語・異なり複合語を数える指標がある：

$d(i,N)$:要素 w_i を取る異なり語数

Nakagawa (2000) はこれを用語抽出に適用し有用性を示している。Baayen (2001) は接辞についてこの指標を「使用の広がり」と呼んでいる。

$d(i,N)$ は、w_i の語レベルでの生産性 (どれだけの複合語を生産し得るか) を反映している。例えば、w_i と w_j について、次のような状況だったら：

	$f(i,N)$	$d(i,N)$
w_i	10,000	100
w_j	1,000	700

w_i よりも w_j の方が造語力すなわち語彙空間における生産力は高いと言えるかも知れない。一方：

	$f(i,N)$	$d(i,N)$
w_i	10,000	1,000
w_j	1,000	700

では、w_i の方が w_j より造語力すなわち語彙空間における生産力が高いとは言いにくい。というのも、w_i は w_j と異なる頻度で現れており、別のテキスト量あるいはもし両者の延べ度数が同じくらいだったら、w_j の取る異なり数の方が大きい可能性が直感的にも明らかだからである。

ここでは造語力の顕現型と潜在性との区別が問題となっている。$d(i,N)$ を顕現型の指標として語彙論的に意義付ける議論は、概略次のようになろう。「テキスト・コーパスがある語彙域の全体を表しているとき、特定共時態における語彙は全て出現している。従って $d(i,N)$ は語彙論的特徴をそのまま表している」。

この議論には問題がある。語彙域に対し包括的なテキスト・コーパスは技術的に入手し得ないとすると、顕現型の議論においてもサンプルたるテキストから「(現実) テキスト域の量」に対して $d(i,N)$ はどう動くのかを巡る予測が必要となる。これは N を現実量として動かした時の $d(i,N)$ の値である。従って、造語力の顕現型の指標は、

$d(i,\lambda N)$ $(\lambda > 0)$

の λ による推移のかたちが表すことになる[4]。

一方、潜在性を意義付ける議論は次のようになる。「文が潜在的に無限であるという主張は延べの語使用が無限であることを意味する。この状態における語彙の構造に見られる要素の「造語力」を本来的な造語力と考えることができる」。潜在造語力 $d(i)$ は、テキスト域依存性を全く除外した場合として定義できる。

$d(i) = d(i,\lambda N)$ $(\lambda \to \infty)$

$d(i,\lambda N)$ と $d(i)$ は、理論上は、二項補間・補外 (Good & Toulmin 1956) 及び LNRE モデル (Baayen 2001) により計算することが可能である。

3 指標と確率分布

$f(i,N)$、$d(i,N)$、$d(i)$ という指標に対応する確率と確率分布を与えることができる。まず、いくつかの分布を整理しよう。

(1) 理論的語彙空間 (単純語と語彙化された複合語を含む辞書)[5] の中で語構成要素 (本稿の検討範囲では主要部) が用いられる確率。事象空間は全ての語構成要素 (主要部) で、確率は要素が取る複合語の比率に対応する。

(2) ある要素がテキスト中で主要部として用いられるときに、その要素を主要部とするある単純語・複合語が選ばれる確率[6]。要素毎に確率分布が定義され、確率は、要素が取る単純語・複合語に対して与えられることになる。

(3) テキスト空間において要素が用いられる確率。事象空間は全ての語構成要素 (主要部) で、確率は要素の出現回数の比率に対応する。

ここで、$f(i,N)$ は (3) の分布、$d(i)$ は (1) の分布における各要素の確率を考えることに対応している。$d(i,N)$ は (3) のテキスト出現確率に各要素における (2) の分布を考慮した指標と言うことができる[7]。

従って、$d(i)$ は、$N \to \infty$ とすることで全ての w_i について $f(i,N) \to \infty$ として (3) の分布を除外しながら、テキスト上に観察される (2) の分布を利用するものである。

4 実験と観察

4.1 データ

NII 言語タグ付きコーパス人工知能分野抄録 (Okada 2001) を利用し (表 1)、いくつかの語構成要素について上記尺度 $f(i,N)$、$d(i,\lambda N)$、$d(i)$ を調べた。

抄録数	延べ語数 (単/複)	異なり語数 (単/複)
1816	299846/230708	8764/23243

表 1　NII コーパス・データ

今回は、「システム」、「モデル」、「知識」、「情報」の 4 要素を調べた。統計的手法が適用できる程度に頻度の高いものから、一般的な要素 (前者) 二つ、分野依存性が高そうな要素 (後者) 二つを選んでいる。表 2 は、この 4 要素について、$f(i,N)$、単純語としての延べ出現数 ($f_s(i,N)$)、複合語主要部としての延べ出現数 ($f_c(i,N)$)、$d(i,N)$ を示したものである。

	$f(i,N)$	$f_s(i,N)$	$f_c(i,N)$	$d(i,N)$
システム	1970	723	1247	502
モデル	1015	328	687	263
知識	1191	748	443	137
情報	637	369	268	155

表 2　4 語の基本データ

[4] 異なる要素の比較を同一テキスト・サイズ λN でではなく、延べ語数 f を揃えて $d(i,f)$ として比較することも考えられる (Kageura 2003)。これは、談話域が中心か語彙域が中心かの解釈上の選択にもよる。例えば、異なる分野の語彙における中心要素を比較するときには、同一の f のもとで比較することが解釈上適切な場合があるように思われる。佐藤 (2003) の、用語を起点とする「マイクロドメイン」概念等も含め、語彙論の特徴の解釈の構図については今後の検討を要する。

[5] 主要部に現れる要素数が有限であり、かつすべての要素についてそれが構成しうる複合語数が有限ならば理論的語彙空間は有限、そうでなければ無限となる。

[6] 検討の対象を主要部に絞っている理由はこの説明にある。つまり、テキストにおいて何らかの概念が選び取られるとき、その中核は概念を表す語の主要部 (類) であると考えるのである。単・複を別のものとするならば、これらは、異なる語の出現頻度に過ぎない。なお、例えば「キノコ分類システム」といった用語が用いられるときには、「システム」を要請するディスコース域と「キノコ分類」を要請する域とが (大きさを違えて) 存在すると考えられるから、この点については今後の検討を要する。例えばテキスト・ウィンドウを柔軟に変えられる指標 (Hisamitsu & Niwa 2000) などを用いて、これを検証するのは有用であろう。

[7] 丁寧に言えば、p_{t_i} を w_i のテキスト出現確率、N をテキスト量、w_i における (2) の分布を w_i を主要部とする単純語・複合語全体からなる標本空間 $S = \{i_1, i_2, i_3, ..., i_{d(i)}\}$ に確率 p_{i_k} が与えられているものとし、二項分布の組み合わせで考えると、$\mathrm{E}[f(i,N)] = p_{t_i} \cdot N$ で、

$$\mathrm{E}[d(i,N)] = \sum_{m=1}^{p_{t_i} \cdot N} \sum_{k=1}^{d(i)} \binom{p_{t_i} \cdot N}{m} p_{i_k}^m (1 - p_{i_k})^{1-m}.$$

4.2 LNRE モデル

これらの要素について、3節 (2) に相当する分布をもとに、$d(i,\lambda N)$ と $d(i)$ を求める。このために、LNRE モデルを利用した（Baayen 2001)。母集団事象として $w_1,...w_S$ の S 個が存在するとし、二項分布（とそのポアソン近似）に従うとすると、大きさ N のデータにおける異なり事象数 $V(N)$ の期待値は、

$$\mathrm{E}[V(N)] = S - \sum_{i=1}^{S}(i-p_i)^N = \sum_{i=1}^{S}(1-e^{-Np_i}).$$

となる。これを、構造異なり分布 $G(p) = \sum_{i=1}^{S} I_{[pi \geq p]}$（ただし I は $p_i \geq p$ のとき 1、それ以外は 0）を用いて書き替えると次のようになる。

$$\mathrm{E}[V(N)] = \int_0^\infty (1-e^{-Np})dG(p).$$

ここで、$G(p)$ に、対数正規分布や一般化逆ガウス・ポアソン分布といった語彙頻度分布の法則を当てはめればよい。ただし、ここで、これらの分布のパラメータは標本量に依存してしまうから、これらの分布が当てはまる最適標本量 Z という概念を導入し、そこにおける分布則の当てはめに基づいて補外を行えばよいというのが、LNRE モデルの基本的なアイディアである。あてはめの妥当性に関する検定の指標も制限された範囲ではあるが利用できる。

4.3 観察と検討

4 要素に対する LNRE モデルのあてはめは表 3 のようになった。$d(i)$ の推定値も共に示す（GIGP は一般化逆ガウスポアソンモデル、IGP は逆ガウスポアソンモデル、LogN は対数正規モデル、MSE は平均二乗誤差)。

	モデル	p 値	MSE	$d(i)$
システム	GIGP	0.96	2.19	273402688337
モデル	IGP	0.47	2.88	3676671255
知識	LogN	0.88	2.72	689
情報	IGP	0.84	2.32	667

表 3 LNRE モデルのあてはめと $d(i)$

データが小さいこともあるかも知れないが、p 値は「モデル」を除いて非常に高い。「モデル」も含めて、当てはめたモデルが妥当性をもっていると考えて良いであろう。

図 2 は、$d(i,\lambda N)$ の値を各モデルに従って、$\lambda = 2$ までプロットしたものである（X 軸はテキスト量)。ここからは、$d(i,N)$ から見てあまり顕著な変化は認められない。全体としては、テキストの増大に対して、「システム」の異なり増加率が高いのに対して「情報」「知識」がかなり低いことがわかる[8] 。

$f(i,N)$、$d(i,\lambda N)$、$d(i)$ の指標で 4 要素を並べると次のようになる。

$f(i,N)$	シ	>	知	>	モ	>	情
$d(i,\lambda N)$	シ	>	モ	>	情	>	知
$d(i)$	シ	>	モ	>	知	>	情

[8] 成長率を取ることもできるが、ここでは視覚に基づく議論にとどめる。

$f(i,N)$ と $d(i,N)$ では「知識」が「モデル」及び「情報」よりも下に来ており、また、$d(i,N)$ と $d(i)$ では、「知識」と「情報」が逆転している。

全体として、各論では、「知識」と「モデル」が、出現頻度としては「知識」が多いにもかかわらず、異なり量としては「モデル」が多くなっている。「知識」と比べると、「モデル」は、テキスト中で用いられるためには諸要素と結びついて使われることが必要な要素であることを示している。また、「知識」と「情報」との関係において、N においては、「情報」が延べ出現が少ない割に異なり語数が多いため、テキスト量が変化しても安定してこの関係が保たれると思われたが、$d(i)$ でみるとわずかではあるが「知識」の方が総異なり語数は多いという結果になっている。

観察した要素が 4 つと少ないため、一般化は難しいが、「システム」、「モデル」などの一般的な要素の方が、「知識」や「情報」などの分野依存性が高い概念を表す要素よりも、主要部として出現する場合、他の要素を取り複合する可能性がより一般に高いのではないかという仮説をたてることができよう。詳細な検討が別途必要になるが、この解釈は、少なくとも「システム」や「モデル」といった要素に対しては当てはまる。ただし、「システム」や「モデル」は実体よりも形式属性を付加する側面があり、どちらかというと概念の性質上特別なものであるという感覚も強い。

専門語彙論の観点からは、これに関わって、「一般的」「分野依存的」といったカテゴリーの検討も必要であろう。今後、他要素の観察や異分野の観察そしてその意味付けも含め、専門語彙論の計量的側面を展開していきたいと考えている。

5 おわりに：語彙空間<の>モデル

「統計的言語モデル」「確率的言語モデル」といった言葉がしばしば用いられ、様々な「統計的言語モデル」を巡る議論がある。これらは何のモデルを議論しているのだろうか。N グラムモデルや HMM は、それ自身としては言語のモデルではない（別の様々なものに応用できる)。とりあえず、コーパスを対象にパラメタを確定して初めて言語のモデルになるならば—そしてこれはコーパスの地位やモデルの一般化域といった問題を抜きにすると正しいと思われるのであるが—、統計あるいは確率のモデルを横断するパラメタの具体値の言語における一般化こそが、本来的な言語のモデルであろうし、言語の研究である限りそうでしかあり得ない[9] 。

「テキスト・コーパスからの語構成要素の語彙論的特徴の復元について」と題する本研究は、そのための方法論的な考察（1 節から 3 節）を中核とするもので

[9] 立川 (1995) は、「言語をめぐる・・・具体的対象のない思考にとどまるのか。それとも、多種多様な諸言語（les langues）という具体的な対象との格闘のために、より整合的で説明力のある分析を求めて言語理論を打ち立てるのか。言語哲学か、言語学か」という問題提起をしているが、それとのアナロジーで形式のモデルか言語のモデルかを問うことは無益ではなかろう。

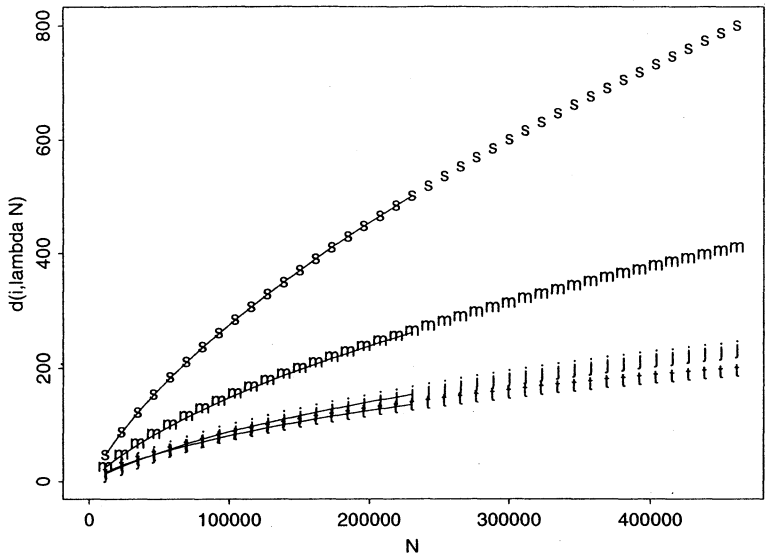

図2 $d(i, \lambda N)$ ($\lambda < 2$) の遷移 (s=システム、m=モデル、t=知識、j=情報：実線は二項補間)

あるが、それは語彙空間のモデル化のための解釈の枠組み及び分析手法の検討と位置づけられる。本来的に、語彙のモデルそのものは、4節で例示したような記述の包括的展開という具体的なレベルに存するのであり、その計算に用いた二項補外やLNREモデルに存するのではない (cf. Kageura 2002) [10] 。

また、本研究は処理を目的とした応用ではないため、精度等の評価がモデルを決めるわけでもない。さらに、語彙論の予測問題が「許容される形式」ではなく「現実的存在可能性」に関わる以上、モデルの妥当性は歴史的軸に関与せざるを得ないため、記述・予測力の評価は今後を待たなくてはならない。

謝辞

本研究は日本学術振興会科学研究費「情報理論的アプローチによる専門語彙構造の特徴解析とモデル化」(基盤 (C)(2)-14580465) の一部として行われた。東京大学情報基盤センターの中川裕志先生、日立中央研究所の久光徹博士には、語彙統計を巡り、京都大学の佐藤理史先生にはドメインと語彙の解釈を巡り有益な議論をさせて頂きました。感謝します。

References

Baayen, R. H. (2001) *Word Frequency Distributions.* Dordrecht: Kluwer.

Foucault, M. (1969) *L'Archéologie du Savoir.* Paris: Gallimard.

Geeraerts, D. (1994) "Lexicology." In: Asher, R. E. *The Encyclopedia of Language and Linguistics.* Vol. 4. Oxford: Pergamon, p. 2189–2192.

Good, I. J. and Toulmin, G. H. (1956) "The number of new species, and the increase in population coverage, when a sample is increased." *Biometrika.* 43(1), p. 45–63.

Hisamitsu, T., Niwa, Y. et. al. (2000) "Extracting terms by a combination of term frequency and a measure of term representativeness." *Terminology.* 6(2), p. 211–232.

Kageura, K. (2003) "On some statistical measures for corpus-based lexicology." *Actas de VIII Simposio Internacional de Communicacion Social.* Santiago de Cuba, 20-24 January. p. 456–459.

Kageura, K. (2002) *The Dynamics of Terminology.* Amsterdam: John Benjamins.

前田富祺 (1989)「語彙総論」玉村文郎編『日本語の語彙と意味 (講座日本語と日本語教育 6)』東京: 明治書院. p. 1–22.

McCarthy, M. J. (1991) "Lexis and lexicology." In: Malmkjaer, K. *The Language Encyclopedia.* London: Routredge, p. 298–305.

Nakagawa, H. (2000) "Automatic term recognition based on statistics of compound nouns." *Terminology.* 6(2), p. 195-210.

Okada, M., et. al. (2001) "Defining principled but practically manageable lexical units in Japanese textual corpora." *NLPRS 2001 Workshop on Langauge Resources in Asia.* Tokyo, 30 November. p. 47–53.

佐藤理史, 佐々木靖弘 (2003)「ウェブを利用した関連用語の自動収集」情報処理学会自然言語処理研究会, NL-153-8.

立川健二 (1995)「解説　ソシュール理論の言語学的転回」フランソワーズ・ガデ著『ソシュール言語学入門』東京: 新曜社. p. 201–214.

[10] それゆえ計算に用いる＜形式の＞モデルがオリジナルなものかどうかは言語の研究の意義には一切関係しない。