日本語専門用語の量的構造の分析について

影浦 峡
学術情報センター研究開発部
kyo@rd.nacsis.ac.jp

要約
日本語専門用語の語種(外来語・漢語)ごとの量的特性を記述するために、与えられた用語データに基づく補間・補外の枠組みを導入し、語彙成長等の指標を定義する。その枠組みを実際の専門用語データに適用し、現実的な範囲での、量的特徴の推移の記述を行う。

1 はじめに

専門用語の造語は分野毎に異なり、また、日本語では外来語とそれ以外(主に漢語)とは異なった役割を果たしていることが指摘されている。この点を巡る記述的・数量的な分析はいくつかあるが(Ishii, 1987; 石井&野村, 1984)、データの静的な記述が中心である。

専門用語の計量的研究を展開するためには、課題が二つある。第一は、専門用語のどのような特徴を記述・予測したいかに応じた適切な量的指標を定義することであり、第二は、専門用語データは一般に小さく、ほとんどの統計的尺度やパラメータは標本の大きさにより系統的に変動するため、それを考慮した計量的枠組みを用いなくてははならないことである。

この点を勘案し、また、専門用語というきわめて具体的な言語現象の計量的特徴付けとしては母集団特性の要約的記述よりも現実的範囲での予測力を重視した方が有用であるという前提のもとに、二項分布に基づくモデルによる補間・補外の理論を応用し、専門用語における語構成要素の特徴付け指標として、延べ・異なり比、及び語構成要素成長の指標を定義する。さらに、実際の専門用語データにこの枠組みを適用し、量的指標を手がかりに、分野間での相違及び分野内での語種による特徴の記述を試みる。

2 専門用語データ

本研究では、専門用語集の見出しデータを用い、語構成要素(語基)の延べと異なりの関係に着目する。テキストでの用語の分布を考慮しないのは、語彙的語構成である専門用語における語基の量的規則性は用語の延べ利用頻度とは無関係であると仮定できるからである(Sager, 1990)。

実際の分析には、計算機科学(相磯, 1993)、農学(文部省, 1986a)、心理学(文部省, 1986b)、物理学(文部省, 1990)という四分野の学術用語データを用いる。基本的な数量を表1に挙げる。ただし、Tは専門用語数、Nは専門用語を構成する語基の延べ数、$V(N)$は、延べ語基数Nにおける異なり語基数である。N/Tは一専門用語の平均構成語基数、$N/V(N)$は一語基の平均使用度数である。これらから、外来語と和漢語の一般的な違いが分野に関わらず認められると同時に、両者の位置づけが分野毎に異なっていることが認められる。

ところで、このようなデータからは、残念ながら、

Domain	T	N	V(N)	N/T	N/V(N)	C_L
計 全	14983	36640	5176	2.45	7.08	0.211
外来語		14696	2809		5.23	0.242
和漢語		21944	2367		9.27	0.174
農 全	15067	29142	9093	1.93	3.20	0.256
外来語		2610	1513		1.73	0.300
和漢語		26532	7580		3.50	0.247
心 全	6272	14314	3594	2.28	3.98	0.235
外来語		1541	995		1.55	0.309
和漢語		12773	2599		4.91	0.207
物 全	10635	25095	4745	2.36	5.29	0.228
外来語		5048	2081		2.43	0.269
和漢語		20047	2664		7.53	0.197

表 1. 標本専門用語データの基本的数量

直接母集団の特性に対する統計量を算出することはできない。というのも、明らかに未出現の語基があるため、標本としては不十分であることが多いからである。例えば、各語基の母集団出現確率推定値に相対頻度を用いることはできない。これは、母集団出現確率を相対頻度で推定して求めた異なり語基数と、実際の異なり語基数との相違を見ると明らかである。この相違は、次節で述べる二項分布を前提とすると、次の式で求められる。

$$C_L = (V(N) - \hat{E}[V(N)])/V(N) = \frac{\sum_{m\geq 1} V(m,N)(1 - p(i_{[f(i,N)=m]},N))^N}{V(N)}$$

ただし、

$f(i,N)$: 語基 w_iの標本 Nにおける出現頻度
$p(i,N) = f(i,N)/N$: 標本相対頻度
m : 出現頻度度数
$V(m,N)$: 標本 Nにおける度数 m の異なり語基数(頻度別異なり語基)
$\alpha(m,N) = V(m,N)/V(N)$.

表1の最後の欄は、それぞれのデータにおける全体及び語種ごとのC_Lの値である。ここから、標本相対頻度を用いると、約2割から3割という大きな誤差を生むことがわかる。語の出現に関わる統計量においては、標本からは信頼できる母集団推定ができない。

3 理論モデルと補間・補外

3.1 理論モデル

母集団において異なり語基がS個 $(w_i, i = 1, 2, ...S)$ あり、それぞれの確率がp_iであると仮定する。語基がランダムに分布しているとすると、二項分布およびそ

れに対するポアソン近似を用いて、大きさ Nの標本における異なり語基数は、次のようになる。

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

$$E[V(m,N)] = \sum_{i=1}^{S}\binom{N}{m}p_i^m(1-p_i)^{N-m} = \sum_{i=1}^{S}(Np_i)^m e^{-Np_i}/m!. \quad (2)$$

以下の準備のために、(1) 式を積分表現で書き換えよう。まず、語彙頻度分布を $G(p) = \sum_{i=1}^{S} I_{[p_i \geq p]}$（ただし $p_i \geq p$ のとき $I=1$、その他の場合は 0）という形の構造異なり分布式で表現する。これを用いると、(1) 式は

$$E[V(N)] = \int_0^\infty (1-e^{-Np})\, dG(p). \quad (3)$$

と表せる。ただし、値の昇順に同じ値の p_iをまとめて添字を付け直した、少なくとも一つの語基が取る確率 p_jにおいて $dG(p) = G(p_j) - G(p_{j+1})$、他の場合は $dG(p) = 0$ である。

3.2 二項補間・補外

ところで、今、大きさ N_0の標本が与えられたときに、それに条件づけられた、大きさ Nの標本における総異なり語基と頻度別異なり語基とは、次のように表すことができる (Efron & Thisted, 1976; Good, 1953; Good & Toulmin, 1956)[1]。

$$E[V_{N_0}(N)] = V(N_0) + \sum_{m=1}^{N_0}(-1)^{m-1}V(m,N_0)\left(\frac{N}{N_0}-1\right)^m$$

$$E[V_{N_0}(m,N)] = \sum_{k\geq m}V(k,N_0)\binom{k}{m}\left(\frac{N}{N_0}\right)^m\left(1-\frac{N}{N_0}\right)^{k-m}$$

いずれの場合も、N/N_0あるいは N/N_0-1 を m 乗するため、残念ながら $N \simeq 2N_0$の周辺から発散し始めるが、与えられた標本の 2 倍程度までの異なり語基についての予測はできる。

3.3 補間・補外の適用とランダム性仮定の妥当性

補間・補外を含む理論的枠組みは、語基の分布に関して二項分布、それゆえランダム性を仮定していた。用語データにはテキストのような順序性はないため (cf. Baayen, 1996)、用語レベルではランダムであると考えることができる[2]。それゆえ、標本の大きさ以内の経験的語彙成長等を見るためには、標本から用語レベルでランダムに複数回抽出した値の平均を取れば良い。

図 1 は、四分野について、標本に基づく二項補間と補外による、$2N_0$までの、全体および語種別の $V(N)$ および $V(1,N)$ の語彙成長曲線と、用語レベルでの 1000 回のランダムパーミュテーションに基づく N_0までの語彙成長の 20 区間のプロットを重ねたものである。N_0までの両者の一致度から、用語レベルでのランダム性を仮定できれば、現実的には、語基レベルでのランダム性を前提とした理論をそのまま用いることができることがわかる[3]。また、$V(1,N)$ の補外については、$2N_0$前で発散していることがうかがえる。

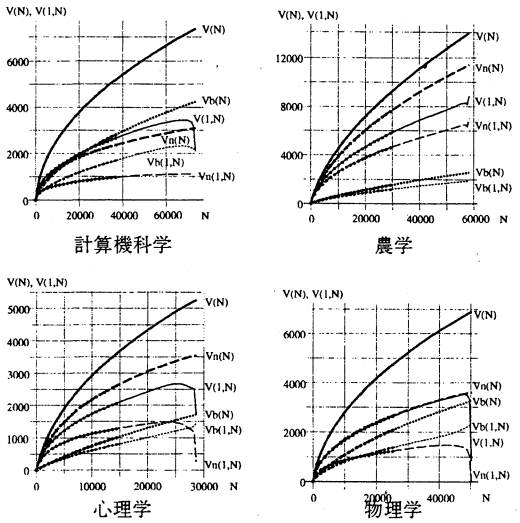

図 1．$V(N)$、$V(1,N)$ の二項補間・補外
（V：全体、Vb：外来語、Vn：和漢語）

4 語基の量的構造と専門用語

前節までで、$N \simeq 2N_0$付近までの異なり語基及び任意の頻度別異なり語基に対する予測の枠組みができ、それが専門用語データに関しては近似的に成り立つことが確認できた。本節では、それを前提に、語基の量的構造の分析を行う。

4.1 量的指標の定義とその解釈

量的指標として、以下に二つを定義する。第一は、語基の平均頻度で、次式で表される。

$$X(N) = \frac{N}{V(N)} \quad (4)$$

我々が対象とするのは、専門用語のレベルでは異なりであるから、語基が複数回出現する場合は、同一語基が同一用語中に現れるというまれな場合以外は、異なる用語に用いられていることになる。従って、用語のレベルから見ると、$X(N)$ の変化は、一つの共通語基により関連づけられる平均用語数を示していると解釈することができる。それゆえ、これを**用語関連度**と呼ぶこととする。

第二の、さらに重要な指標として、成長率 $P(N)$ を定義する。(1) あるいは (3) 式を用いて異なる Nに対

[1] 超幾何分布に基づくべきだが、二項分布で近似する。

[2] ただし、「中心的用語」といった語彙の階層を考えると、ある種の語彙の分類が必要になるが、ここではそれは考えない（cf. 榎沢, 辻&影浦, 1998)。

[3] $N_0/2$ までについては t-検定を容易に行える。結果に有為な差はなかった。

するプロットを取ると、語彙成長曲線が得られる。(3)式を微分すると、以下のような、Nにおける**成長率**が得られる。

$$P(N) = \frac{d}{dN}E[V(N)] = \frac{E[(V(1,N)]}{N} \quad (5)$$

これは、大きさNの時点で、さらに標本を増やしたときに新しい語基が現れる確率を示している（Baayen, 1991）[4]。

語種との関係で$P(N)$を考えた場合、語種毎に算出する$P_i(N)$と全体において算出する$P_f(N)$とを定義できる。例えば、外来語について、以下のようになる。

$$P_{fb}(N) = E[V_{borrowed}(1,N)]/N$$

$$P_{ib}(N) = E[V_{borrowed}(1,N)]/N_{borrowed}$$

$P_i(N)$においては、もともとの標本専門用語集合はそこから語種毎の標本を取るための元データに過ぎず、従って、$P_i(N)$は、語種別々に見たときの成長率を示すものである。これに対して$P_f(N)$は、大きさNの専門用語集合において実際に新語基が現れる確率を示しており、語種別にみるならば、いわば、新概念要素を担う語基の語種毎の特性を示していると解釈できる。$P_f(N)$を特に**新規性率**と呼ぼう。

4.2 用語関連度の推移

まず、用語関連度から見てみよう。図2は、四分野について、全体、外来語、和漢語の用語関連度を、全体の標本Nを横軸にして示したものである。

まず、全体の指標から、農学の用語だけ、極端に低い値で用語関連度が推移していることがわかる。物理学と心理学は、観察できる範囲では似通っており、計算機科学はそれより少し高い値で推移している。語種に注目すると、全ての分野で和漢語よりも外来語の用語関連度が低い値で推移していることが観察できる。心理学と農学では、外来語の平均仕様度数は似通っており、非常に低い。一方、計算機科学では、かなり高い値で推移している。語基の役割という観点からは、一般に外来語が特異的な概念要素に個別に割り当てられ、用語間の関連づけという用語集合における体系表示機能を担いにくいと解釈できるが、計算機科学においては、平均でみる限り全体の用語体系に馴染んだ使われかたがされていると考えられる。和語については、計算機科学と物理学の値が似通ったところを推移しており、心理学がそれより多少低い値で推移する。農学における値は全体同様、他分野と比べて極端に低い。

分野別に傾向をまとめると、計算機科学では、そもそも一語基あたりの用語関連度が高く、既に現れた語基を新たな造語に繰り返し用いる傾向が強い。その中で、和漢語の使用度数が高いが、多分野との比較によると、外来語もかなり使用度数が高くなっている。農学は反対に全体の用語関連度がそもそも非常に低く、相対的にはその値の低さは、和漢語の用語関連度の低さによっている。物理は、和語の用語関連度は計算機科学と同様であるが、外来語の用語関連度が低い。心理では、外来語の用語関連度が農学と同様低いが、和語の関連度は計算機科学や物理学に近いため、全体の傾向は物理・計算機科学に近くなっている[5]。

[4]ちなみに、これはGood-Turing推定による、標本非出現語全体の確率と等しい。

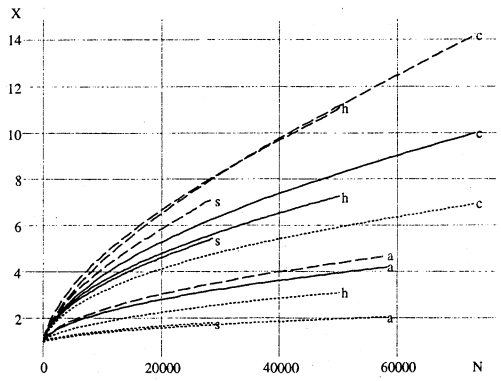

図2．$X(N)$の推移
（c:計算機・a:農・s:心理・h:物理）
（実線:全体・破線:外来語・ダッシュ:和漢語）

4.3 新規性率の推移

図3は、四分野における新規性率の推移を、それぞれ$2N_0$まで示したものである。

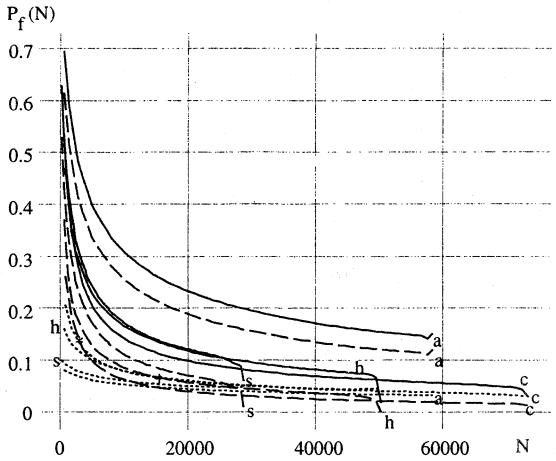

図3．$P_f(N)$の推移
（c:計算機・a:農・s:心理・h:物理）
（実線:全体・破線:外来語・ダッシュ:和漢語）

分野毎にみると、全体においては、用語関連性と対象的に、農学における新規性率が一貫して高い値で推移することがわかる。これに対し、計算機科学は初期において新規性率が急速に低下し、また、心理学と物理学では、計算機科学よりも急ではないがそれに近い推移を示す。

[5]$V(N)$が$N \to \infty$において有限だとすると、当然、$N \to \infty$において$X(N) \to \infty$であり、ここから逆に、与えられた標本の絶対的な大きさが、用語関連度の推移にある程度影響する疑いが生じる。これについては、用語の中心性・周辺性の問題とも関わって、今回は扱えなかったので、今後の課題としたい。

語種毎にみると、Nが小さいうちは、各分野とも、和漢語の新規性率が高いが、Nが大きくなるにつれ、一般に和漢語の新規性率は急速に低下し、新規性率における外来語の比重が高まってくることがうかがえる。実際、計算機科学と物理学では、それぞれ$N \simeq 3500$及び16000で両者の値が逆転している（図中*および+）。用語関連度では語種の開きが大きかった心理学でも、両者の値は急速に接近していっており、Nをさらに増やせば、両者の関係が逆転することが予測される。

表1を参考に、これらを検討すると、特に興味深いのは、延べ語基の比率で外来語が全体の4割近くを占める計算機科学や全体の2.5割を占める物理学では、既に与えられた標本の範囲内で外来語と和漢語の新規性率が逆転していることが観察されるが、一方全体の1割程度である心理学についても、現実的なNの範囲内で恐らく両者が逆転しそうであると予測される点である。これに対して、心理学と延べの外来語の比率がそれほど変わらない農学では、Nを相当増やさない限り、両者の関係は逆転しそうにない。これは、農学においては、むしろ和漢語が他分野における外来語的機能をはたし続けているからと考えられる。

これらを考えると、外来語は用語関連度が低く、一方新規性率がNが大きくなるにつれて増大する、すなわち、新規概念に個別に新外来語表現を当てることが、用語が増大するに従って多くなる、といった、語種毎の一般的特性は明らかに存在するが、個別分野の造語特徴と語彙の量的構造とを詳しく検討するためには、語種という分類をカテゴリカルに前提するのではなく、同一語種の語基の分割等も含めて詳しい検討が求められることがわかる。個別の要素に着目した分布特性の検討やそれに基づく分類が必要となってくる。

5 おわりに

本研究では、$N \simeq 2N_0$程度の大きさまで補間と補外を用いて語基成長を予測し、用語関連度と新規性率を指標として専門用語における語種の役割の推移を記述してきた。これにより、分野毎・語種毎の特徴を、動的な予測も含めて記述することができた。これは、専門用語の語彙論的な理論として重要であるのみでなく、テキスト中に現れうる語種と用語との関係についての指針を与えるため、自動専門用語抽出のような応用にも利用可能であろう。

補外範囲については、標本データの大きさにもよるが、現実的存在可能性を問う場合でも、ここで扱った$2N_0$を越える範囲への補外について扱えるように手法を展開する必要がある。これに関して、Good & Toulmin (1957) は、発散を抑える技術的提案を行っている。一方、Chitashvili & Baayen (1993) は、語彙頻度分布として明示的な理論分布を仮定し、それを組み込んで任意の大きさへの補外・母集団語彙量予測を行う理論的枠組みを展開している。

補外の拡張と関連して、具体的なNの値が、専門用語においてどのように位置づけられるのかを考えなくてはならない。これは、与えられた標本データの母集団との関係における位置づけの問題に関連しており、用語の中心性・周辺性の問題とも関わって、理論的には重要な課題である。

語彙的特性の指標と関連しては、本分析では、純粋に語基の集合としての性質を見たにとどまる。これに対して、複合語の構造上の位置や結合特性をも考慮した分析を、場合によっては別の指標も検討しつつ進めることが必要となろう。これらについては、これまでに述べてきた問題点・解釈の可能性の追求とともに、今後の課題としたい。

補記

本研究における考えを整理するにあたって、マックスプランク心理言語学研究所の Harald Baayen 博士に多くの示唆をいただきました。同氏に感謝致します。

References

相磯秀夫 (編) (1993) 情報処理用語大辞典. オーム社.

Baayen, R. H. (1991) "Quantitative aspects of morphological productivity." In: Booij, G. and van Marle, J. (eds.) *Yearbook of Morphology 1991*. Dordrecht: Kluwer. p. 109–149.

Baayen, R. H. (1996) "The randomness assumption in word frequency statistics." In : Perissinotto, G. (ed.) *Research in Humanities Computing 5*. Oxford: Oxford University Press. p. 17–31.

Chitashvili, R. J. and Baayen, R. H. (1993) "Word frequency distributions." In: Hrebicek, L. and Altmann, G. (eds.) *Quantitative Text Analysis*. Trier: Wissenschaftlicher Verlag. p. 54–135.

Efron, B. and Thisted, R. (1976) "Estimating the number of unseen species: How many words did Shakespeare know?" *Biometrika*. 63(3), p. 435–447.

榎沢康子, 辻慶太, 影浦峡 (1998) "専門用語コーパスにおける語彙的な階層付けの可能性." 言語処理学会第4回年次大会 (1998年3月24日–26日).

Good, I. J. (1953) "The population frequencies of species and the estimation of population parameters." *Biometrika*. 40(3–4), p. 237–264.

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.

Ishii, M. (1987) "Economy in Japanese scientific terminology." In: Czap, H. and Galinski, C. (eds.) *Terminology and Knowledge Engineering*. Frankfurt: Indeks Verlag. p. 123–136.

石井正彦, 野村雅昭 (1984) "機械工学用語の語種構造." 計量国語学. 14(4), p. 163–175.

文部省 (編) (1986a) 学術用語集：農学編. 学術振興会.

文部省 (編) (1986b) 学術用語集：心理学編. 学術振興会.

文部省 (編) (1990) 学術用語集：物理学編. 第2版. 培風館.

Sager, J. C. (1990) *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins.