データを生み出す知識、知識を生み出すデータ
― 「正しい」コーパス認知言語学のための覚書 ―

吉川 正人　(慶應義塾大学 [非常勤])

1. はじめに

近年、PC の発展・普及伴い、電子コーパスを利用した言語研究はますます盛んになっている。特にオンラインで手軽に大規模コーパスを検索することのできるサービスの登場およびその普及の影響は大きい。[1)]黎明期は内省ベースの研究が中心であった認知言語学も、近年では盛んにコーパスデータを利用した分析が取り入れられており、特に「用法基盤モデル (Usage-based Model: e.g., Kemmer & Barlow 2000)」を冠した研究ではもはや主流となっていると言っても過言ではない。

このような状況において懸念されるのは、「コーパスの乱用」とでも言うべき事態である。コーパスはあくまで単なるデータの羅列であり、そこから一定の手法でデータを抽出・分析し知見を得るには、相応の手続きが必要であるが、アクセス・検索が容易であればあるほど、この手続きの部分には目が行きにくくなる。本稿では、このような問題意識から、認知言語学においてコーパスを用いる際、特に言語知識の実態解明のためのツールとしてコーパスを用いる際に、自覚的になっておくべき点を整理する。具体的には、1) 代表性・均衡性というコーパスの持つ一般的性質が言語知識の探求においてどう影響してくるかを確認した上で、2) コーパスには「産出データのサンプル」と「入力データのサンプル」という二つの側面があり、それぞれどのような研究のアプローチがあり得るか、またどのような限界があるかを、研究事例を紹介しつつ議論する。また最後に、3) コーパスデータと内省データの比較の元、コーパスによって何が分かり、何が分からないのか、という点について議論する。

2. コーパスの性質と言語知識

一般的に、コーパスを評価する基準として「均衡性 (balancedness)」と「代表性 (representativeness)」という二つの性質が取り上げられることが多い。均衡性とは、コーパス内のデータ構成として特定のジャンル・レジスターや年代への偏りの程度をさす。偏りがなければないほど「均衡」である、ということになる。もちろん、「話しことばコーパス」や「雑誌記事コーパス」のような、特定のジャンルやレジスターに特化したコーパスは存在するが、その場合は、その特定の範囲内での偏り度合いを可能な限り低くすることが求められる。代表性とは、コーパスデータが何らかの母集団のサンプルであると見做した際に、その母集団の性質をどれだけ反映したものであるか、という指標である。現代アメリカ英語のコーパスであれば、そのコーパス上のデータが現代アメリカ英語の様々な側面を満遍なく抽出していることが望ましい。

通常、均衡性が高ければ代表性も高くなるが、これは均衡性の定義によるところも大きい。例えば、書き言葉・話し言葉双方を満遍なく収録した大規模均衡コーパスにおいては、多くの場合「新聞」「雑誌」「フィクション」など複数のジャンルから同程度の分量のデータがサンプリングされるが、[2)] このような配分は必ずしも代表性を最大化しているとは考えられない。この点に関しては後述する。

2.1. コーパスと言語知識

コーパスを用いて言語知識の探求を行うということは、コーパスデータが何らかの形でヒトの言語知識の有り様を反映している、と考えるからである。次節で詳述するように、この考えには二種類の想定があり、一つは、

コーパス上のデータの分布はヒトの「言語行動」の記録であり、その行動を生み出しているヒトの知識を反映したものである、という考え方であり、もう一つは、コーパス上のデータの分布がヒトの「言語経験」の記録であり、その経験が言語知識を形作っている、という考え方である。次節で両側面における可能性と限界について詳しく見ていくが、本節ではひとまず、この区別にかかわらず、一般的に均衡性・代表性が言語知識の解明にどのような影響を与えるかを考える。

2.2. コーパスの均衡性と言語知識

均衡性について問題となり得るのは、コーパス上のデータ構成が、ヒトの言語行動・言語経験の有り様と対応していない場合である。例えば、はたしてコーパス上のレジスターやジャンルの構成比が、ヒトが実際に入力として得ているジャンル構成比と対応しているか、ということは問題になり得る。

多くの大規模均衡コーパスは書き言葉が大半で、話し言葉の割合は 1-2 割程度であることが多い。[3)]しかし、たった数分の談話であっても文字に起こすと相当の長さとなることを鑑みると、実際我々が普段触れている言語情報としては、個人差もあるが、少なくとも 5 割程度、場合によってはそれ以上は話し言葉であろう。従って、書き言葉を中心とした均衡コーパスを、言語知識の解明のためのツールとして使用するのは、均衡性の観点から考えるとやや問題を孕むものかもしれない。

これに関連して、Landauer & Dumais (1997) は、コーパスデータから語の意味を自動獲得する「潜在意味分析 (Latent Semantic Analysis, LSA)」という手法の紹介にあたって、以下の事実を指摘している: アメリカの典型的な中学一年生 (7 年生) は 1 日に新たに 10-15 単語を学習しているが、1) その大多数が書き言葉にしか登場しないこと、2) 話し言葉で遭遇する単語はほぼ全て既知であること、3) 直接教示を受けて学習する語は平均 1 語以下であることから、語の学習は基本的には書き言葉をベースに行われる (Landauer & Dumais 1997: 211)。このことが意味するのは、少なくともヒトが持つ語の知識をコーパスベースで解明していく場合には、書き言葉を中心とするコーパスを用いる方がむしろ好ましい、ということである。

このように、コーパスの均衡性、或いはデータのレジスター・ジャンル構成比率に関しては、言語知識のどのような側面を解明したいのかに応じて、適切性を柔軟に判断する必要がある。そのような目的に応じて、場合によってはコーパス全体を検索対象とするのではなく、一部のレジスター・ジャンルに限定する、といった対処が必要となろう。

2.3. コーパスの代表性と言語知識

代表性に関しては、Taylor (2012) が 3 つの観点から問題提起を行っている。1 点目は、本節冒頭で述べた、均衡性との関連で生じる問題である。世論調査や市場調査をする場合、世論や市場のニーズを把握するには全人口の男女比率や世代間の人口比率を反映した形で対象者をサンプリングすることになるが、同様の「構成比率の反映」を言語データに対して達成するのは困難である (Taylor 2012: 13-14)。

2 点目は言語データの受容者の問題である。上記の問題を解決する一つの手段として、一人の個人が数週間、数年、あるいは一生涯など、一定期間内に受容した言語データを全て集める、という方法が考えられるが、その場合も、例えば当人の職業によって得ている言語経験は劇的に変わりうるわけで、その「個人」としてどのような人物を選定するか、という問題がつきまとう。結局、データの代表性の問題が、選定する人物の代表性の問題にすり替わったのみで、問題の解決には至らない。[4)]

3 点目は特定の言語表現に関する問題である。ほぼどんなテクストにも登場し、従ってどんな人物であっても触れる機会のある表現もあれば、一部のテクストにしか登場せず、従って一部の人物にしか接触機会の無い表現というものもある。この「一部のテクスト」をコーパスの一部に組み込んでしまった場合、代表性は著しく低下することとなるが、あらゆる表現に対してそのようなテクスト間の分布の偏りを見積もることは極めて困難であろう。実際、以下のような事例が報告されている。Stefanowitsch & Gries (2003) はコーパス上で特定の構文とそ

の構文に生起する動詞の共起頻度から構文と動詞の「相性」を統計的に計算し容認性を予測する「共起構文分析 (Collostructional Analysis)」という手法を考案し、その適用例として International Corpus of English のイギリス英語版 (ICE-GB) を用いた命令文の共起構文分析を行っている。その分析において、命令文との共起強度第 6 位に fold という動詞が検出されていたが、実際は fold を用いた命令文のすべてが折り紙の折り方マニュアルという単一のテクストに生起するものであった (Taylor 2012: 15)。[5)]

以上のように、言語知識探求のためにコーパスを用いる場合、代表性の観点から複数の問題が指摘できる。分析に当たっては、これらの問題を考慮し、コーパス上の頻度を絶対視せず、必要に応じて何らかの補正を加える、といった対応が必要になろう。

3. コーパスの 2 側面

前節で述べたように、コーパスには「人々が実際に使った言語データのサンプル」としてのコーパスと、「人々が実際に見聞きした言語データのサンプル」としてのコーパスという 2 つの側面が存在する。前者の側面を「産出データのサンプル」としてのコーパス、後者を「入力データのサンプル」としてのコーパスと呼ぶことにする (Table 1)。コーパスを用いて言語知識の探求を行う際には、この差異について自覚的になり、自身がそのどちらの側面に着目しているかを認識すべきである。本節では、それぞれの側面に着目しコーパスを分析している研究事例を紹介し、両者の差異と可能性について考察する。

3.1. 「産出データのサンプル」としてのコーパス

コーパスデータはいわば言語の「使用実態」のサンプルであり、「人々がどのように言語を使用しているか」を映し出すものと言っていいだろう。例えばコーパス上で、ある一定の条件下、あるいは文脈で一定の言語使用が行われる、という事実が確認できれば、それは言語行動のパターンの記述になり、その言語行動の背後に何らかの言語知識を想定することが可能となろう。

例えば、筆者は以前、幼児の縦断発話データ (Brown 1973) から、ある時点での統語知識の状態を推定し、月齢ごとの変化を捉えることで、統語発達のプロセスを計算的に示す、という研究を行っていた (吉川 2010, 2011, 2012)。幼児がある時点で give me some と発話したとして、この幼児がどのような言語知識を持っているかをこの発話単体から推定するのは容易ではない。give me some という総体を一語文 (Holophrases) として記憶しており、それをそのまま使用しているのみである、という可能性もあるし、あるいは give me X というスロットを含むパターンの実現として想定することもできるだろう。もしくは、[Verb Object_1 Object_2] という、抽象的な統語パターン (項構造構文) の実現として使用している、と想定することも可能かもしれない。このような多様な可能性の中から適切な記述を選択する方法として、筆者は、当該の幼児の過去の発話データを参照する、という方法を用いた。つまり、その幼児が "give me one" "give me orange juice" など、give me X の実現例と思われる複数の発話を産出していたとしたら、 give me X というパターンを知識として持っていると想定可能であろう。このように、過去の発話の集積から、部分的に一致するパターン (e.g., give me X) を網羅的に見つけ出し、そのパターンの変項の実現値 (e.g., some, one, orange juice) にバリエーションがある場合、そのパターンを幼児の持つ言語知識として認定する、という方法を採用した。

この手法は、幼児の「発話」という目に見える行動データから、その行動を可能にしている知識状態を逆算して推定する、という試みであると言える。ただしこの手法には一つ大きな限界が存在する。通常幼児個人の縦断データをとる場合、1 か月や数週間といった感覚で自宅を訪問し、1 時間程度周囲の大人との会話を収録する、という方法を用いる。従って、コーパスデータ上には収録されていない期間の言語行動は一切現れない。収録期間の言語行動が非収録期間も含めた当該幼児の全言語行動を偏りなく反映していれば問題はさほど大きくないかもしれないが、「偏りの無さ」を保証するのは困難で

Table. 1　コーパスの持つ二つの側面

1.	コーパス as 産出データのサンプル	言語行動	言語知識の反映
2.	コーパス as 入力データのサンプル	言語経験	言語知識の源 (形成要因)

ある。これは前節で指摘した代表性の問題である。筆者の分析は同一幼児の縦断データから、発達に伴う統語知識の変化という相対的な情報を抽出することを試みたものであり、そこに一定の傾向が見て取れれば何らかの発達プロセスを補足したとみなして問題はないかもしれないが、使われている言語表現それ自体など、絶対的な情報を分析対象とする際には細心の注意が必要である。

3.2. 「入力データのサンプル」としてのコーパス

多くのコーパスベースの認知言語学的分析はコーパスを入力データのサンプルとみなす研究であると言える。この背景には、ヒトの言語知識が何らかの形で入力として得ている言語経験の統計的性質を反映している (Cf. 分布バイアス仮説: Shirai & Andersen 1995) という前提がある。これは所謂「用法基盤モデル (Usage-based Model: e.g., Kemmer & Barlow 2000)」の基本的な想定であろう。代表的な研究としては Joan Bybee による一連の研究 (e.g., Bybee 1995, 2010) が挙げられる。Bybee (1995) ではヒトのもつ形態論的な知識が接触頻度を反映した (表示の)「強度 (lexical strength)」を伴って蓄積されているとする表示モデルを提案し、コーパス上の語や形態素のタイプ・トークン頻度でその強度を近似している。また、言語習得の領域では同種の研究には枚挙に暇がないが、例えば Goldberg, Casenhiser, & Sethuraman (2004) では、幼児と周囲の大人の会話を収録したコーパス (Bates, Bretherton, & Snyder 1988) を用い、大人から子供に向けられた発話に見る頻度分布から、構文の習得にはその構文に生起する動詞の頻度分布が「偏った (skewed)」ものであることが重要であると言う結論を導き出している。また筆者も、「一定の環境化で頻繁に共起する表現は一つのユニットとして認識・学習される」という想定の下、英語の前置詞与格構文 (e.g., John gave a book to me.) の分析を行い、コーパス (British Nationl Corpus) 上の動詞と前置詞 to の共起強度が有意に高いことを示している (Yoshikawa 2008)。

このアプローチの最大の問題は、コーパスデータの統計的な分析を、ヒトが言語学習に際して実際に行っている言語データの統計的な処理と同一視する必要がある、ということである。コーパスデータとヒトの得ている言語入力が対応づけられているということは、前者の統計的な分析は後者の統計的な処理であり、前者の分析結果は後者の処理の結果としての言語知識ということになる。ヒトが言語入力に対して実際にどのような統計処理を行いどのように学習を行っているのか、ということが依然として未知であることを鑑みると、これは大いに問題であろう。関連して、2.3 で言及した共起構文分析に対して、Schmid & Küchenhoff (2013) はそこで用いられている統計指標の適切性を問題視する議論を紹介し言語知識の表示との関連で考察を加えている。今後もこのような議論を絶やさず、どのような統計指標をどのような場合に用いるべきか、指針作りを進める、といった対応が必要となろう。[6)]

4. 結語

以上見てきたように、コーパス上のデータを用いて言語知識の有り様を探求する際に、均衡性・代表性の観点から適切にコーパスやコーパスの一部を選定するか、分析結果を補正する必要があり、また、分析にあたってコーパスをヒトが産出したデータのサンプルと見做しているのか、入力として得ているデータのサンプルと見做しているのかについて自覚的になり、そのいずれかに応じて適切なデータの選定・処理を行う必要がある。

本節では、最後に、もう少し俯瞰的な観点から、コーパスデータを言語知識の探求に用

いることの意義・位置づけを考える。コーパスと対照されることの多い方法として、内省を用いた言語分析が挙げられる。また、内省で得られた記述が、コーパス上のデータの分布と合致しない、ということも多々報告される。その場合、内省/コーパスいずれを重視する立場かに応じて、いずれか一方のデータが正しく、もう一方は何らかのノイズの影響で「歪んだ」データであると見做されることがほとんどであろう。しかし実際は、前者は無意識の言語行動に反映される行動パターンの実現であり、後者は意識的な言語に対するメタ認知に反映される規範意識の実現である、と考えることも可能であり、いずれか一方が「誤っている」と考える必要はないかもしれない (吉川 2015)。

筆者は、前者はヒトの個体が接触経験から得たものであるという点で極めて「個人的」(或いは「個人認知的」) な性質を持つものである一方、後者は個々人がメタ認知の際に推定する、所属する言語コミュニティで通用する規範であるという点で極めて「社会的」(或いは「社会認知的」) であると考えている。この特徴づけが正しいかどうかは経験的に検証する必要があるが、いずれにせよ、両者の間には言語知識の性質として大きな差異がある可能性が指摘できる。今後はこの差異についてより議論・考察を深め、両データから得られた一般化を補完的なものとして統合していく試みが必要となってくると考えられる (Table 2)。

例えばコーパスデータから得られた一般化を内省によって検証する、或いはその逆に内省によって得られた一般化をコーパスデータで検証する、という手続きをとる際には、両者の差異について意識的になっておくべきであろうし、より一般的には、それぞれの性質に基づいて内省データを用いるべき対象とコーパスデータを用いるべき対象を適切に判断する必要があるだろう。

恐らくコーパスに現れる言語行動のパターンは、レジスターやジャンルといった領域毎に大きく異なるものであり、当該言語全体に共通する一般化を行うことは言語知識の実態を反映していないと思われる。

注

[1)]代表例は Brigum Young 大学の Mark Davies 氏による複数のコーパスとその検索インターフェースであろう (http://corpus.byu.edu/)。

[2)]例えば現代アメリカ英語の均衡コーパスである *Corpus of Contemporary American English* (COCA, Davies 2008-) では、「雑誌」「新聞」「フィクション」「学術書」「話し言葉」がそれぞれ 20% ずつ配分されている (http://corpus.byu.edu/coca/help/texts.asp)。

[3)]脚注 2) で述べたように、COCA では 2 割程度である。イギリス英語の大規模均衡コーパスである *British National Corpus* (BNC, http://www.natcorp.ox.ac.uk/) においては、書き言葉が 9 割程度で、話し言葉は 1 割ほどしか収録されてない (http://www.natcorp.ox.ac.uk/corpus/)。

[4)]この点に関しては、個人の一日の会話行動を、世代・性別の偏りなく 200 人ほどから収集し、日常会話の均衡コーパスを構築しようという試みがある (小磯ほか 2015)。これは日常会話というジャンルに限定したコーパスであるが、同様の調査を、調査スパンを長くし、書き言葉も含めて行うことも原理的には可能であろう。

[5)]このようなテクスト間の偏りを考慮し、Gries (2008) では「分散 (dispersion)」という指標を用いて頻度を補正する、という手法を提案している。

[6)]一つの方向性として、行動実験とコーパスデータの統計的分析結果の対応を見る、というアプローチがあり得る (e.g., Gries, Hampe, & Schönefeld 2005)。

参考文献

Bates, E., Bretherton, I., & Snyder, L. S. 1988. *From first words to grammar: Individual differences and dissociable mechanisms.* Cambridge: Cambridge University Press.

Brown, R. 1973. *A first language: The early stages.* Cambridge, MA.: Harvard University Press.

Bybee, J. 1995. Regular morphology and the lexicon. *Language and cognitive processes*, *10*(5), 425–455.

Bybee, J. 2010. *Language, usage and cognition.* Cambridge: Cambridge University Press.

Davies, M. 2008-. The Corpus of Contemporary American English (COCA): 400+ million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. 2004. Learning argument structure generalizations. *Cognitive Linguistics*, *15*(3), 289–316.

Table. 2　分析対象と使用データの対応例

	分析対象	前提	データ
1	規範意識としての言語の実態	規範はメタ意識に現れる	内省
2	集団としてのヒトの言語行動 (における統計的性質)	言語は統計的・社会的実態である	コーパス (as output)
3	個人のもつ言語意識	意識は知識を純粋に反映する	内省
4	個人の言語学習の結果	言語学習は入力データに依存する	コーパス (as input)

Gries, S. T. 2008. Dispersions and adjusted frequencies in corpora. *International journal of corpus linguistics*, *13*(4), 403–437.

Gries, S. T., Hampe, B., & Schönefeld, D. 2005. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, *16*(4), 635–676.

Kemmer, S., & Barlow, M. 2000. Introduction: A usage-based conception of language. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language* (pp. vii–xxviii). Stanford: CSLI Publications.

小磯花絵・土屋智行・渡部涼子・横森大輔・相澤正夫・伝康晴. 2015. 均衡会話コーパス設計のための一日の会話行動に関する調査–中間報告–. 第 7 回コーパス日本語学ワークショップ予稿集 (pp. 27–34).

Landauer, T. K., & Dumais, S. T. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Schmid, H.-J., & Küchenhoff, H. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics*, *24*(3), 531–577.

Shirai, Y., & Andersen, R. W. 1995. The acquisition of tense-aspect morphology: A prototype account. *Language*, *71*, 743–762.

Stefanowitsch, A., & Gries, S. T. 2003. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, *8*(2), 209–243.

Taylor, J. R. 2012. *The mental corpus: How language is represented in the mind.* Oxford: Oxford University Press.

Yoshikawa, M. 2008. How can we find syntactic patterns?: A case study of English PDC as an embodiment of usage-based view of language. *Colloquia*, *29*, 95–104.

吉川正人. 2010. パターンの生産性に見る統語発達: パターン束モデルに基づく習得プロセスの検証. 日本認知科学会第 27 回大会発表論文集 (pp. 235–241).

吉川正人. 2011. 生産性の漸増が語る統語知識の発達: パターン束モデルに基づく段階的発達プロセスの計算的実証. 日本認知言語学会論文集 (第 11 巻, pp. 618–621). 日本認知言語学会.

吉川正人. 2012. スキーマの計算理論を求めて: 漸進する統語発達過程の記述問題とその解法. 認知言語学論考 (第 10 巻, pp. 193–246). 東京: ひつじ書房.

吉川正人. 2015. 文法性判断の社会言語学: 社会統語論の目論見. 社会言語科学会第 36 回大会発表論文集 (pp. 26–29).