日 本 独 文 学 会 研 究 叢 書
「コーパスをめぐって　心理・知覚表現の分析」　田中　愼 編　所収（印刷中 2009）Version 1

方法としてのコーパス？ ― 言語研究におけるデータの扱い

吉田 光演　（広島大学）

0．はじめに ― 内省 vs. コーパス ―

ドイツ語など自然言語の研究のアプローチには様々の立場があるが,言語研究を科学的な営みにしたいという思いは共通していると思う。確かに,「科学的とは何か」の理解において既に見解が分かれるが，少なくとも対象（言語）に関するデータの抽出と観察に基づいた客観的（間主観的）記述が共通項となることについて異論はないだろう。理論言語学では，さらに記述を超えて抽象的モデル化や仮説・検証による説明といった方法も問題となる。いずれにせよ，経験的言語データに基づかずに「ドイツ語は○○だ」などと主張することは反証不可能であるが故にすべきでない。一定の現象について一定の方法で抽出したデータに対して一定の方法で分析すれば，類似した結果が再現できるということが言語の科学でも求められる。言語データの収集としては，調査（フィールド，アンケート），実験（内省，言語実験による反応測定）などと並んで，言語資料（コーパス）の分析が考えられる。データ抽出の方法は，何を研究目的とするかということと不可分であり，特定の方法が絶対というわけではない。過ぎ去った遠い過去の時代の歴史的な言語状況を分析する際には，当然ながら実験的方法は使えず，言語資料も制限されている場合が多い。現代ドイツ語の話し言葉や社会言語学的現象に焦点を当てる時には，会話例を収集できればよいが,人員･予算の問題や倫理的問題が生じ,多くのデータを集めることは困難である。実例に基づく言語データを多数収集したい場合，今日ではデジタル化された電子コーパスを使うのが便利である。コーパスによる実例分析は，作例によるデータの文法性判断という方法（内省）へのアンチテーゼとして対立的に強調されることが多い。たとえば，生成文法のように，文法モデルに基づく言語能力のあり方を記述･説明する目的で例文を作成し,その文法性判断を話者の判断に委ねる内省的方法は，不自然で恣意的なものに陥りやすいという批判である。こうした問題意識からのコーパスへの傾斜は二重の意味で危険である。一つは,「内省（人為）vs. 実例（自然)」という対立軸自体が誤っている。ま

た，内省判断と実例分析とは相容れないものではなく，両立しうる。生命科学などでは，in vitro （試験管）実験と，in vivo （生体内）観察でできることとできないことがあり，目的と状況に応じて使い分けるのが当然であり，両方行う場合もある。これと同様のことが内省とコーパスにも言える。

この小論では以下，方法としてのコーパスの問題点について考察し，また，今回のシンポジウムで議論された論点について筆者の見解を述べたい。

１．コーパスは言語共同体の発話の総体たりうるか？

本論集田中論文で Bloomfield が引用されているが，「ある言語共同体において用いられている発話の総体」をコーパスとして収集することは実際には不可能である。コンピュータの処理能力が向上したこと，インターネット経由のデータ量が爆発的に増えたことを根拠に機械可読コーパスを「言語共同体において用いられた発話の総体」と等値することはできない。理由は単純で，人間が産出する文の数は制限がなく，無限生成される。他方その時々に蓄積されるコーパスは有限だからである。確かに IDS の COSMAS のようなコーパスデータは膨大で，小説や新聞雑誌等の書き言葉コーパスとしての価値は大きい。しかし，Textsorte をすべて網羅しているわけではない。たとえ，（Brown Corpus などを参照して）多様なテクスト種を平均的に抽出し，テキスト量を一定化してサンプル化した形の「バランス（大規模均衡）コーパス」を作成したとしても，それがドイツ語共同体の発話総体という母集団から抽出した標本であるという保証はない。Newmeyer (2005: 160f.)が指摘するように，英語の代表として *New York Times* が挙げられるかもしれないが，英語母語話者の子供が *New York Times* を読んで英語を習得するわけではない。特定コーパスが言語共同体全体の言語・文法を反映しているということは論理的にありえない。むしろ，言語共同体の発話データのモデルの一つとしてコーパスを相対化する視点が必要である。相対化によって，話し言葉と書き言葉の相違，テクスト種間のスタイルの差異など，サンプルとしてコーパスの意義が浮かび上がるだろう。

２．コロケーション分析の背後にある言語直観

もちろん上述の指摘によってコーパスの価値が薄れるわけではない。1000 万語以上の大規模コーパスには多数の母語話者の言語使用の傾向がある程度反映されており，それを統計的に推測し，確率論的推定を行うことは可能である。Bloomfield のもう一つの主張である「分布主義」の今日的形態としてのコロケ

ーションも，コーパスを対象として一定の傾向を抽出することは可能であり，辞書記述の際の用例記述には役立つ。また，外国語としてのドイツ語学習において，語彙から句レベルの表現を作る際に役立つ。しかし，言語研究者全員が辞書学の専門家になるのではないから，コーパスが言語研究の一部を占めることはあっても，その中心になることはない。高頻度コロケーションを知ることは重要だが，エントロピーで見れば，あまりに高頻度のものは陳腐すぎて情報価が下がる(ich bin…の変化など）。高頻度語彙にしても上位 100～200 語では意義は低い。ライプチヒ大学 Wortschatz プロジェクトによるドイツ語 Top 100 のリストでは，der, und, sich, in など機能語，前置詞，代名詞がほとんどで，内容語は Prozent, Mark, Jahr, wieder, Uhr, immer, Millionen, sagte くらいである(http://wortschatz.uni-leipzig.de/html/wliste.html)。逆に確率が低すぎる表現も利用価は低い。比較的高頻度で，情報性の高いコロケーションを選択すること，言語運用以外の要因による過剰な出現をフィルタリングする必要がある。しかし，この種の操作は「コーパス＝自然」という前提から離れて，コーパスを操作するという人為的視点が介在することを意味している。

話者は，ドイツ語非母語話者である我々を含めて，有限リソースとして蓄積されたコーパスから語とそのコロケーションを常に参照して文を発話するのではない。話者は，言語能力と認知能力に基づいて，（共時的に）有限のレキシコンから語を抽出しながら，個別言語に特有の統語規則を適用して創造的に句と文を生成していく。そうでなければ,「言語共同体の発話総体」はある一点で収束し，言語使用は凍結した「言語法典」を参照することによる「無限の引用」となるだろう（幸いそんな事態は起こり得ない)。コロケーション分析から導き出せるのは，確率的に非常にありそうな結合から，滅多にありそうにない結合に至るまでの相対的･連続的分布であり,それが即客観的解釈を示すのではない。ある特定のコロケーション分布から，その背後にある言語直観・パターンを導き出すのは，それを分析する研究者である。研究者の側にそれなりのデータを振り分ける主体的な視点がなければ，生のデータはいかようにも解釈可能なものでしかない。このことは自然科学分野では当然のことである。実験・観測は一見誰が行っても全く同じ結果が得られる客観的営みであるかのように思われる。しかし，物理学や化学の研究者に聞けば分かるように，実験研究では巧みな技術が必要であり，地道なトレーニングと，何をどのように測るのかという背後の問題意識が必要である。観測・実験を何百回続けたら自然に法則性が見えて，科学的真理に到達するということは経験科学ではありえない。

３．内省（実験）とコーパス分析は矛盾しない

1950 年～60 年代初期のコンピュータ科学の言語解析モデル・形式言語理論の構築に大きな影響を与えたチョムスキーが，その派生物であるコーパスデータの統計利用について，言語学者として懐疑的見解を表明しているのは，経験科学に関する素朴な誤解への警鐘としても理解できよう。「ワインバーグが言及したガリレオ流というのは，いま構築している抽象的な体系こそが実在の真実であるという認識なのです。現実は，あまりにも多くの要因，あらゆる種類の事柄のために,真実がなんらかの形で歪曲されたものなのです。だとすると，すべてに注意を払うことはできないということを認識しながらも，現象自体は無視して，ある事物はなぜそうあるのか，ということに実際に深い洞察を与えるだろうと思われる原理を追求することは，しばしば意味のあることだと思えるのです。物理学者たちは，たとえば今日でさえ，水がどのように蛇口から流れ出すのか,あるいは,ヘリウムの構造や他の複雑すぎるように思えること等，詳細には説明できません。」（チョムスキー 2002）

ここで，コロケーション分析・言語統計が,「蛇口から出る水の流れ」の記述と同じレベルだという意図はない。「ここにある言語使用のデータこそが客観的現実であり，言語研究の真の対象である」といった主張に対しては，チョムスキーの指摘は妥当な批判たりうるという意味である。さらに言えば，特定のコロケーションを設定する段階で，既に一定の「原理的な問題設定」が持ち込まれている（それを意識するか否かの問題は別として）。たとえば,清野論文の「コーパス準拠型研究」は，まさに原理的な問題設定に基づく検証の手段としてコーパスが選ばれる。他方，特定の問題意識に基づく研究の中で，思いがけない発見がなされたケースについて，清野は「コーパス駆動型」研究であると述べているが，そのような研究の副産物は他のアプローチでも十分に生じることである。それを発見するかどうか，それをどのように料理するかは，研究者の研ぎ澄まされた問題意識やテクニックに依存するのであり,「コーパスが（研究を）駆動している」わけではない。

データへの過度の依存に対する警告は，生成文法理論のみならず，他の言語研究者からも発信されているものであり，とりたてて特別なことではない。

> 内省に依存する場合に比べて，コーパスからはおいしい料理とまずい料理が同時に大量に出される。それをどう食べるかは，ひとえに研究者がそこから何を栄養と

して吸収したいかという研究課題にかかっている。Halliday (1996)が言うように，コーパスが勝手に文法記述をしてくれることは決してない。(深谷 (1998), 142)

> the most important skill is not to be able to program a computer or even to manipulate available software (...) [but] to be able to ask insightful questions which address real issues and problems in theoretical, descriptive and applied language studies. (Kennedy 1998: 3).
>
> The use of both introspection and corpus-based analysis can contribute to linguistic analysis and description. Corpora cannot tell us everything how a language works. For example, they cannot be used as a basis for stating what structure or processes are not possible… The fact that an item or structure does not appear in even the largest corpus does not necessarily mean that it cannot occur, but could suggest that the corpus might be inadequate or the item infrequent. (Kennedy 1998: 8)
>
> …the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body. (C. Fillmore 1992:35) (コーパスデータからの経験的研究と内省・理論研究の両方が必要という意味)

引用したように，コーパス研究を主導する研究者もコーパスのみが分析の駆動力となることはないことを強調する。Kennedy の指摘は，(保阪論文でも述べられているが)，「非文から見えるものがコーパスからは見えない」ことを認めている。意味機能文法の大家 Halliday や，格文法の提案者・構文文法の立役者である Fillmore なども，言語データの観察の意義を認めつつも，それだけで研究が完結するわけではない点に（当然ながら）注意を喚起している。

問題は内省実験の側にもある。「文法的・非文法的」な例文の内省による判断は様々の要因が絡む可能性があり，意味的に容認不可能だが文法的な文，意味的に想定可能だが非文法的な文など，単純な 2 項対立で処理しきれないケースも多々ある。理論に有利な例だけを列挙する方法では肝心の言語能力を測定できない。実験デザインについての検討や個人差などの要因の考察などの問題意識が問われるのである（北川・上山 2004)。ドイツ語場合，Fanselow (2004) でも触れているように，例文の文法判断（*の有無)，容認度の判断（?, ??，あるいはその中間）の微妙な相違から，北ドイツと南ドイツの方言の違いによる統語的な相違（主語と目的語の抜き出しの可能性など）だけでなく，個々の集

団の文法意識の相違も見えてくる（Fanselow 2004: 485f.）。

さらに内省観察だけでなく，コーパスから実例を抽出できれば，仮説検証・反証の支えになる（内省とコーパスは背反しない）。たとえば，weil 文などの副詞節では定動詞 2 位（V2）も可能な場合があるが，「dass /ob 節などの動詞の項としての補文では V2 は不可能」という主張が生成文法でも繰り返しなされてきた（筆者自身これを現代ドイツ語文法の（規範的）規則と考えてきた）。しかし，Freywald (2008)は，Freiburger Korpus などコーパスによる分析をも利用しつつ，現代ドイツ語会話では dass 節で V2 が生じることを観察・例証した（"ich weiß, dass herr LAACK hat eine STIFtung gegründet. (ARD, Talkshow)）。Freywald によれば，14300 個の dass 節のうち，V2 の dass 節が 50 個あった（0.34%）。頻度としては小さく，その文法的性質の評価は今後の課題となる（枠外配置による見かけの V2 の問題，補文標識 dass の文法的特性などの理論的な考察が不可欠）。しかしこの例は，理論と内省だけでは見えてこない言語観察とコーパス分析による驚きの発見であると言える。

４．心理動詞のコーパス観察から見えてくるもの

上述の批判は本論集の他の論者に直接向けたものではない。各論文の分析は，コーパスを使うといっても，個別のコロケーションの内側まで踏み込んだ分析であり，記述の対象がコーパスから取られたものという条件がつくだけで，他は変わらないからだ。個々の内容について若干コメントを行う。

（清野論文）

清野の心理動詞のコロケーション分析で取り上げられた，interessieren などの他動詞的用法，sich interessieren などの再帰用法，interessiert sein などの受動的な形容詞的用法といった構文の選択には，分析の際に暗黙の仮定が前提される。即ち，動詞が特定の項構造をとること（Valenz 風にいえば，異なる Aktant を支配），項には特定の意味役割（Agens, Patiens, Stimulus）が結びつくこと，統語形式が変化すればこれらの意味関係も変わりうること，有生の人間が主語位置にある場合には，経験者役割がより agentivisch に解釈できる―といった仮定である。これらの仮定は決してトリビアルなものではない。たとえば同じ心理表現でも，単純に Thema が一つ現れる表現（X ist (für mich) interessant）はなぜ取り扱わないのか？（Web 検索すると，X ist mir interessant も若干ヒット）。また，mögen (Ich mag diese Musik) vs. gefallen

(Diese Musik gefällt mir) のような項の生起は，項構造の違い（意味の違い）か，単なる経験者と心理刺激（原因）のコード化（位置の問題）の違いなのか？心理動詞の主語位置は動作主解釈か，原因(Causer)解釈か？背後にある問題意識によって，データの解釈も変化する（量子力学における観察者と観察対象の変化のように)。そこにも研究者の主観的（主体的）判断が潜んでいる。また，清野論文が取り上げた ärgern, enttäuschen, freuen, interessieren, überraschen といった心理動詞の選択は，(i)他動詞的（使役的 Causative)，(ii) 再帰化可能（使役からの論理的可能性：使役→脱使役：自発または経験者主導)，(iii) 状態への降格（(ii)から event 変化を除去）—のプロセスを原理的に含む意味で興味深い選択である。構造化すれば次のようになる：

(1) [A(Stimulus) CAUSE [BECOME [B(Exp) HAVE PSYCH A(Target)]]]
刺激 A が引き起こす（経験者 B が目標 A に心理状態をもつようになる）

興味深いことに，清野が対象としたコーパスによれば，interessieren だけがこの３つの交替をほぼ均等にもち，他の動詞は(iii)の状態をもたないものがあり，また，再帰化を許さないものもあるという。freuen は他動詞形が少ないが，これは他のコーパスでも同じ結果が出るのか知りたい点である（だとすればなぜ？）（es hat mich gefreut, ..zu…など Web 検索では多数ヒットするが)。enttäuschen, überraschen に再帰形が少ないのは, 経験者のコントロール度の少なさが関係するのか，アスペクトも関わるのか（enttäuschen, überraschen は瞬時変化・完結的，ärgern, freuen, interessieren は漸増的，プロセス的なニュアンスがあるかもしれない)。今後の研究が期待されるが，要はこうした実例分析は背後の言語学的問題意識なしでは成立しないという点である。

（保阪論文）

保阪論文では，versuchen の補部選択の揺れが問題になる。versuchen は zu 不定詞補文を選択し，補文の意味上の主語を義務的にコントロールすると言われている。しかし，コーパスからはこの標準から逸脱して，dass 節が選ばれてくる現象が見られるという。その場合，単に dass 節に拡大するのではなく，補文の見えない主語は動作主ではない（むしろ Theme/Patiens)役割でなければならないという制限（仮説）を立てる。

(2) a. *Ich versuche, dass ich möglichst früh aufstehe.
b. *??Ich versuche, dass ich von der Frau einen Blumenstrauß kriege.

しかし，筆者が Web (Google) 検索エンジンでドイツの Web サイトを検索してみたところ，dass 節補部が 800 例程度ヒットした。その中でかなりたくさんあったのが，以下のような書き間違いの例である（ ... zu Infinitiv のところを，dass を付けて，dass ..zu Infinitiv にしてしまった)。

(3) Ich versuche, dass Training so abwechslungsreich wie möglich zu gestalten.

このような例はふるい落とさねばならない（この種の間違いがよくある)。しかし，保阪の指摘にあるように，**Ich will, dass** aus Freundschaft Liebe wird.といった will dass との類推に似た形で，versuchen が dass 補文を取る例が見つかる。そこには，以下の(4)のように補文の主語は nicht-agentivisch な役割のものがある。しかしまた，agentivisch な解釈もできるような(5)のケース（あるいは境界例）もいくつか見つかる（受動化可能)。

(4) Ich versuche, dass ich natürlich ein guten Job bekomme.
Ich versuche, dass wir in der Fastenzeit mit Freunden zusammen sind.
Ich versuche, dass das eine Ausnahme bleibt.
(5) ich versuche, dass du Folgendes verstehst.
ich versuche, dass ich eine geschmeidige Kommunikation zu ihnen entwickeln kann. Ich will dass du mich liebst.

上述のコーパスの種類の制約もあり，また，Webの場合は地域方言差，個人差，文体差など，多くの要因が絡んでおり，統計的・定量的な分析･評価は正確には出せない。保阪のいうように,「ただし､細かく見ていくと相当怪しい例があり、その点では信頼の置けるインフォーマントや､研究者の判断が不可欠であろう」という注意が必要だろう。しかし，仮説提示，仮説とインフォーマントテストでは見過ごしていた微細な問題が見えてくることもある。先に見たFreywaldの補文dassのV2と同様に，従来の定説を覆す発見ができる可能性もある。

結論的には，一定の問題意識のもとにコーパス分析を行うことは言語研究を

進める上で確かに有用である。しかし，生の事実がそこにあるかどうかも問題であり，あったとしても事実を処理する力が必要である。ドイツ語研究を科学的に営み，学生や院生に研究の方法を教える場合，その一つの方法としてコーパスの扱い方を教えることは有益である。しかし，そこから興味深い考察を得て興味深い問題設定（コーパス検索）を行うには，ドイツ語についての直観を鍛えると同時に,先行研究の検討と理論的な仮説設定･方法論的問題意識を研ぎ澄ますことも同等に大切である。科学的探究に近道はないのである。

【参考文献】

Bloomfield, Leonard (1933): *Language*. New York: Holt.

チョムスキー，ノーム（2008）:「自然と言語」（大石・豊島訳），研究社.

Fanselow, Gisbert (2004): Fakten, Fakten, Fakten! *Linguistische Berichte* 200, 481-492.

Freywald, Urlike (2008): Zur Syntax und Funktion von dass-Sätzen mit Verbzweitstellung. *Deutsche Sprache* 2008/3, 246-283.

深谷輝彦（1998）:「コーパスに基づく文法研究」，斉藤俊雄，中村純作，赤野一郎(編)「英語コーパス言語学」，研究社， 123-143.

Halliday, Michael A.K. (1996): On Grammar and Grammarians. R. Hasan et al. (eds.): *Functional Descriptions*. Amsterdam: Benjamins, 1-38.

Kennedy, Graeme (1998): *An Introduction to Corpus Linguistics*. London/ New York: Longman

Newmeyer, Friederick, J. (2005): *Possible and Probable Languages*. Oxford: Oxford Univ. Press.