コーパスの類型論

後藤　斉　（東北大学文学部）
gothit@sal.tohoku.ac.jp

1　はじめに

本論においては，従来とは違った観点からのコーパスの類型論を提案する．これによって，「コーパス」という用語の多義性をより的確にとらえ，「コーパス」の概念をより明確にすることを目的とする．

「コーパス」という用語は，後藤(1995)で指摘したように多義であり，少なくとも広義（「大規模なテキストの集積」）と狭義（「ある言語(の部分集合)を代表すべく集められた大規模なテキストの集積」）とを区別する必要がある．（以下において，特に断りのないところでは「コーパス」を広義で用いる．）

Computer corpora are, essentially, bodies of natural language material (whole texts, samples from texts, or sometimes just unconnected sentences), which are stored in machine-readable form. ... It should be added that computer corpora are rarely haphazard collections of textual material: they are generally assembled with particular purposes in mind, and are often assumed to be (informally speaking) representative of some language or text type. (Leech & Fligelstone 1992: 115-116)

この多義性をうまくとらえるためには，コーパスを適切に分類し，典型的な類と非典型的な類を認めることが必要であろう．その際には，「意図，想定」といった，人間の側の要因を考慮に入れなければならない．

また，現在のコーパス言語学がコンピュータなしには不可能であり，「コーパス」が第一義的には機械可読のものをさすことが当然であるにしても，コンピュータを使わないで大規模データを扱うことも不可能ではない．コンピュータが言語研究に利用されるようになる以前の研究（C.C.Fries 1940, 1952; West 1953)や，現在でも，コンピュータを使わずにコーパス言語学と類似の手法をとる研究(山田 1996a, b)との連続性を無視しえない．したがって，コーパスの定義や分類は電子化の有無とは独立に考える．

2　従来のコーパスの分類

これまで，コーパスはその内容や形態から分類されてきた．すなわち：

1:　竹沢・末松(1995)
　　内容（異種/同種/体系的/専門的）
　　形態（生/タグつき/分析ずみ）

2:　松本・小磯(1996)
　　対象分野（書き言葉/話し言葉，テキスト/音声コーパス...）
　　偏在性　（網羅性，結束性，推敲の程度...）
　　加工度　（マークアップの程度，タグ情報...）

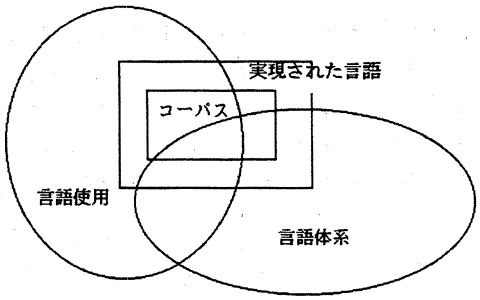
これらの分類は，コーパスそのもの，すなわち集められた結果のみに着目している．したがって，「意図，想定」といった人間的な要因を含むことが難しい．

3 新しい類型論

本発表では，コーパスをその外部との関係に基づいて分類する新しい類型論を提案する．外部とは，コーパスを部分集合とする全体集合である．言い換えれば，コーパスを材料として研究されるべき対象となる言語である．このことをAarts (1991)を参考にして図示すれば，図1の通り．基本的には，コーパスの作成者がコーパスの外部に何を想定していたかによって分類する．すなわち：

(1) テキストそのものを志向するコーパス
　　全体集合はコーパスと同一
(2) 実現された言語のサンプルとしてのコーパス
　　全体集合はコーパスより大きい有限集合
(3) 言語使用の近似物としてのコーパス
　　全体集合は無限集合
　(3a) 単一変種のみに焦点をあてるもの
　(3b) 多変種を同時に扱うもの
(4) 言語体系のデータとしてのコーパス
　　全体集合は無限集合

図1 コーパスとその外部



3.1 テキストそのものを志向するコーパス

コーパス中のテキストそのものが研究の目的である場合である．これは「コーパス」の語源であるラテン語corpusの原義(「体」)からの派生義としての「資料の総体，集積」にもっとも近い．しかし，有限のテキストが目的であるから，一般化より個別の事象に向かいやすく，言語研究より文学研究に近づく．したがって，現代的には非典型的なコーパスである．

> 例：Corpus Inscriptionum Latinarumなどのコンピュータ以前のテキストの集成やOxford Text Archive，勉誠データベースなどの文学作品テキストデータ，CD-毎日新聞などの新聞記事テキストデータ．

3.2 実現された言語のサンプルとしてのコーパス

過去の一定期間にある形態で実現された(例えば，出版された)言語は，有限であるが，膨大であって，直接には扱いにくいことがある．そのような有限の言語を目的としつつ，そのサンプルとして選ばれたテキスト群としてのコーパスがここに入る．

例：The Century of Prose Corpus, Hansard Corpus

C.C. Fries の研究もここに入れることができる．アメリカ構造主義言語学ではコーパスの延長上に言語があるとみなされた．

3.3 言語使用の近似物としてのコーパス

言語使用は容認可能な文ないし発話の集合とみなせるが，これは無限集合であり，その全体を直接に扱うことはできない．それを有限のテキストの集積であるコーパスによって近似することによって，記述・説明を可能にする．大量であるほど近似がよくなることが期待できるので，大規模テキストデータというコーパスの利点をもっとも発揮できる．したがって，これがもっとも典型的なコーパスである．

質的に近似をよくする通常の手段は，一旦，言語使用の中で実現された有限の言語を考え，それにできるだけ相似な部分集合をとることである．この点では(2)に類似する．

言語使用は必然的に変種を含んでいるため，ここからさらに下位分類できる．

(a) 特定の変種のみに焦点をあてるもの

例：CHILDES, The Bergen Corpus of London Teenager Language

(b) 変種の構成に配慮したもの

例：Brown Corpus, British National Corpus

3.4 言語体系のデータとしてのコーパス

言語体系は文法的文の集合とみなせるが，これも無限集合であって，その全部を列挙することはできない．言語体系は言語使用よりも抽象度が高いため，データはその量より当面の議論に関与的であるかの方が重要であって，大規模さというコーパスの特徴を生かしにくい．特に生成文法においては，データとして母語話者の直感が重視される．しかし，理論にとってクルーシャルな例文を集めたコーパスがありえる．また，言語体系が言語使用とは別であるにしても言語使用からまったく独立しているわけではないのであるから，目的によってはデータとして利用することができる．

例：筑波コーパス（原口 1982）

4 作成者の意図と利用者の意図

上ではコーパスの作成意図によって例を挙げた．しかし，コーパス作成者の意図(作成目的)と利用者の意図(利用目的)とが違うこともありうる．例えば，「CD-毎日新聞」は特定の年に発行された新聞の本文そのものを収録しているが，それを現代日本語の(書き言葉の)使用実態の近似物として使うことがある．この場合，「(1)として作成されたが，(3b)の代用として利用する」とみなせる．すなわち，同一のコーパスであっても，その外

部との関係のとらえ方の違いによって分類が異なりうる.

5 おわりに

上でコーパスの新しい類型論を提案した. これは, コーパスそれ自体の内容や形態に基づく分類に取って代わろうとするものではないが, それとならんで有効な分類であると考える.

参考文献

後藤 斉 1995 「言語研究のためのデータとしてのコーパスの概念について」『東北大学言語学論集』4:71-87.

竹沢寿幸, 末松博 1995 「音声・テキストコーパスとその構築技術, 標準化動向」『人工知能学会誌』10:168-180.

原口庄輔 1982 「新しいデータ処理方式を求めて」『月刊言語』11:9:109-115.

松本祐治, 小磯花絵 1996 「日本語のコーパス」『月刊言語』25:10:114-120.

山田忠雄 1996a 『私の語誌 1 他山の石』三省堂.

山田忠雄 1996b 『私の語誌 2 私のこだわり』三省堂.

Aarts, J., 1991 "Intuition-based and Observation-based Grammars", In: K. Aijmer et al. (eds.), *English Corpus Linguistics*. London: Longman. pp.44-62.

Fries, C.C., 1940 *American English Grammar*. New York: Appleton-Century-Crofts.

Fries, C.C., 1952 *The Structure of English*. New York: Harcourt, Brace and Co.

Leech, G. & S. Fligelstone, 1992 "Computers and Corpus Analysis", In: C. S. Butler (ed.), *Computers and Written Texts*. Oxford: Blackwell. pp.115-140.

West, M., 1953 *A General Service List of English Words*. London: Longman.