

### 1.3.5. The corpus approach

2019/11/26 英語学特殊講義 森田真登

言語学は経験科学(empirical science)であり、分析に用いるデータの集め方には、伝統的に三つの方法がある。つまり、内省(introspection)、インフォーマントから答えを引き出すこと(eliciting answers from informants)、テキストを研究すること、である。歴史言語学においては、インフォーマントはずいぶん昔に亡くなっており、それゆえ現存するテキストこそが証人である。コーパスを使用するということは、歴史言語学においても、膨大な量の言語データを抜粋し分析する最も実りある方法の一つとして確立している。コーパスの方法論のいくつかの利点は以下のように言える。コーパスは自然に生じたデータ(naturally occurring data)から為り、そのデータは観察可能で証明可能である。一方で、内省の判断は主観的で観察できない。デジタルコーパスにおいて、言語構造はコンピューターによって検索でき、検索に当てはまるすべての用例を取り出し、頻度を計算し、統計を出すことができる。それによって、言語分析に対してより体系的なアプローチを保証するのである。コーパスは通常大量のデータからなっている為、コーパスによって導き出した結論は、より統計的に有意なものである。コーパスは通常、作品の代表的なサンプルを含んでいる。

たしかに言語調査においてコーパスを使用することは明らかに利点があるものの、コーパスの限界というものも認識しておくことは重要である。Curzan and Palmer 以下のように述べている。「量的調査は文脈や解釈なしではほとんど役に立たない。『数え終わるのが終わると、調査が始まる(Research begins where counting ends.)』言い換えると、コーパスデータは内省と調査の洞察力によって分析、解釈されなければならない。コーパスは通常、特定の期間の歴史的テキストをすべて含んでいるわけではないので、言語学者は調査を行っている特定のデータベース・ジャンルから度を越えて推測してしまってはならない。なぜなら、それらは一般的で全体的に言語を表すというよりも、あくまで特定のディスコースやジャンル内に限った言語を伝えているにすぎないからである。」初期 ME を分析する際は、現存する証拠がまばらである為、こうしたことを心に留めておくことが重要である。McEnery と Wilson は Curzan and Palmer の意見を共有している。「コーパスと内省に基づいた言語学へのアプローチというのは、お互いに相いれないものではなく、相補的であるとみなせれる。コーパス言語学は内省と観察の統合であるし、そうであるべきである。」

抽象的なレベルでは、「英語史のコーパス」というのは「すべての現存する英語史テキストが合わさったもの」と定義できるかもしれない。具体的なレベルでは、スキャンされ時にタグ付けされた英語史テキストを含むかなりの量のコーパスがある。Kyto and Rissanen はより広義の「コーパス」の定義を認めており、「歴史言語学者は、調査する現象の言語的証拠を生み出すテキストやテキストの集合という古い意味でのコーパスや、その電子バージョンという新しい意味でのコーパスに、頼っている。」と述べている。今回用いる LAEME と LALME は電子コーパスで、SMED は、コンピューターで可読ではないものの、こちらもコーパスであることに変わりはない。