

The Penn-Helsinki Parsed Corpus of Early Modern English を検索して

齊 藤 俊 雄

筆者は「初期近代英語における目的語付き動名詞構文の発達—The Penn-Helsinki Parsed Corpus of Early Modern English を検索して—」なる小論を、『英語史研究会会報』(6 月 13 日)の新刊紹介欄で紹介された田島松二(編)『ことばの楽しみ—東西の文化を越えて』(南雲堂、2006 年 3 月刊)に寄稿した。本稿では、PPCEME を検索した同論文について感じている 2 つの点を取り上げる。自分の論文を自分で review するようなことであるが、ご了承願いたい。

1. コーパスの規模と代表性 (representativeness)

周知のように、The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME) は、Pennsylvania 大学のチームが Helsinki Corpus の EModE の部分 (約 60 万語) を 3 倍 (約 180 万語) に拡張したもので、この PPCEME は (1) Helsinki, (2) Penn1, (3) Penn2 の 3 つの subcorpora (各約 60 万語) より構成されている。(1) は Helsinki Corpus の EModE の部分を使い、(2) と (3) は Pennsylvania 大学側が Helsinki Corpus を補足した部分であり、原則的には Helsinki Corpus と同じ資料から未採録部分を採り、不足分は新たな資料から補足している。時代区分やテキストジャンルも Helsinki Corpus をそのまま踏襲している。

筆者は Helsinki Corpus 公開後間もなく「初期近代英語における動名詞の発達—The Helsinki Corpus of English Texts を検索して—」(『近代英語研究』編集委員会編『近代英語の諸相』英潮社、1993, 353–375) を発表している。昨年 PPCME2 と PPCEME を収録した Penn-Helsinki Parsed Corpora of Historical English (CD-ROM) が配布されたのを機会に、規模が3倍のコーパスに拡張されて、どのような検索結果が出るのか興味を持ち、PPCEME を使って EModE の目的語付き動名詞構文の発達の再調査を試み、その結果を上記の論集に寄稿したわけである。

PPCEME を検索した結果、目的語付き動名詞構文 (*the book's reading*; *(the) book reading*; *reading of the book*; *the reading of the book*; *reading the book*; *the reading the book*) の用例数は、次の表の通りであった (カッコ内は1万語あたりの生起率)。

	Helsinki	Penn1	Penn2	総計
E1 (1500-1569)	382 (19.4)	356 (18.3)	322 (17.4)	1060 (18.4)
E2 (1570-1639)	430 (21.9)	489 (21.9)	557 (23.9)	1476 (22.6)
E3 (1640-1710)	572 (31.9)	710 (35.9)	623 (33.2)	1905 (33.7)

この検索結果について、筆者は「Helsinki Corpus におけるこの構文の生起率が、3倍に拡大した PPCEME の生起率とほぼ同じである点は、注目に値することである。＜小規模の＞Helsinki Corpus でも、このような生起率の比較的高い文法項目の調査には有効である証拠であろう」と書いた。動名詞構文の6つの型の分布を示す Table 3 (ここに再録はしない) の統計でも「Helsinki Corpus における型の分布状況は、概して PPCEME 全体の型の分布と平行している」のである。

この調査結果から、動名詞構文のように頻度の高い文法項目の調

査では、苦勞して語数の多い PPCEME 全体を検索する必要がなく、約 60 万語の subcorpus のどれか 1 つを検索すれば十分であることが証明されたと言えよう。

しかしながら、筆者はこの結果について多少の危惧の念を抱いている。PPCEME は、上述のように Helsinki Corpus とほとんど同じ資料を使って拡張されたために、Helsinki Corpus と同質の汎用コーパスになり、このような類似の結果になったのではないのか、全く違う資料を使っていれば、違った結果が出る可能性があるのではないのか、ということである。いわゆるコーパスの代表性 (representativeness) の問題である。規模が 3 倍になったので、EModE をより正確に代表している (represent) として、その検索結果が EModE の実態をより正確に示していると単純に受け止めてよいのか。かりに Helsinki Corpus の資料選択に偏りがあるとすれば、同じ資料で拡張された PPCEME は、その偏りを量的に増幅させただけではないのか。この疑問の検証は、全く別の資料を使った EModE の汎用コーパスができるまで待たなければならないわけである。

英語史研究会第 16 回大会では Helsinki Corpus の拡張版、PPCME2 と PPCEME を使った研究発表があるので、この点についても議論があることを期待している。

2. コーパスとテキストジャンル

筆者は同論文において *the reading of the book* と *the reading the book* という動名詞構文の 2 つの型の競合についての F. Th. Visser (*An Historical Syntax of the English Language*, Part II, Leiden, 1966, §1124) の次の主張を検証している。

... from the beginning of the fourteenth century to the end of the nineteenth century constructions with *of* and those without *of* before the complement of a form in *-ing* preceded by *the* were used side by side—after 1500 with almost equal frequency—...

PPCEME 全体を検索した結果は、次の表の通りである。

	the reading of the book	the reading the book	合計
E1	331 (94.6)	19 (5.4)	350 (100%)
E2	443 (91.3)	42 (8.7)	485 (100%)
E3	276 (62.3)	167 (37.7)	443 (100%)

筆者は「これだけの用例数が得られれば、 χ^2 検定をするまでもなく、1500 年以降についての Visser の主張は妥当性を欠くと判断できるであろう」と断定した。なお「1500 年以降については」は「1500 年以降の頻度については」の舌足らずの表現である。この表の統計を見ると、確かに、E1 (1500-1569), E2 (1570-1639) では *the reading the book* 型は未発達であり、Visser のいう ‘after 1500’ は妥当ではないであろう。その点は、M. Tajima (*The Syntactic Development of the Gerund in Middle English*, Tokyo, 1985, p. 79) が late ME でのこの型の未発達の状況から Visser の主張を疑問視していることを裏書している。

筆者の軽率さは、Visser の主張を安易に否定し去ったことである。E3 (1640-1710) でのこの型の急増にもっと注意を払って追求すべきであった。

汎用コーパスの強みは、まさに、さまざまなテキストジャンルのテキストを包含していることである。PPCEME は 16 のテキストジャンルを持っている。ジャンル別による両型の頻度を調べると、E3 では次ページの表の通りになる。*the reading the book* 型の頻度の百分率の降順に並べられている。

この統計表から、各ジャンルの用例数は少ないが、ジャンルによっては、*the reading the book* 型が *the reading of the book* 型を凌駕する、あるいは拮抗する頻度を示していることが分かり、上掲の Visser の主張に続く “and that the choice between the two patterns was largely, if not exclusively, a question of style.” という発言に納得させられる。こ

のジャンルによる分類でも、E1, E2 では *the reading the book* 型は未発達で、*the reading of the book* 型を数量的に凌駕する、あるいは拮抗する用例数を示すジャンルはない。従って、Visser の主張は、‘after 1500’を ‘after the mid-seventeenth century’ と修正すれば、一応両構文の並立する状況の時期の上限を示すことになるであろう。（もちろん PPCEME が EModE を代表しているという前提に立っての話ではあるが。）

	the reading of the book	the reading the book	Total
Biography, other	8 (38.1)	13 (61.9)	21 (100%)
History	12 (40.0)	18 (60.0)	30 (100%)
Handbook, other	12 (52.2)	11 (47.8)	23 (100%)
Travelogue	7 (53.8)	6 (46.2)	13 (100%)
Letters, non-private	18 (54.5)	15 (45.5)	33 (100%)
Law	83 (56.1)	65 (43.9)	148 (100%)
Letters, private	14 (63.6)	8 (36.4)	22 (100%)
Science, other	32 (74.4)	11 (25.6)	43 (100%)
Diary, private	22 (75.9)	7 (24.1)	29 (100%)
Educational treatise	30 (78.9)	8 (21.1)	38 (100%)
Proceedings, trials	4 (80.0)	1 (20.0)	5 (100%)
Sermon	18 (81.8)	4 (18.2)	22 (100%)
Philosophy	7 (100)	0 (0)	7 (100%)
Drama, comedy	4 (100)	0 (0)	4 (100%)
Fiction	4 (100)	0 (0)	4 (100%)
Biography, autobiography	1 (100)	0 (0)	1 (100%)
Total	276 (62.3)	167 (37.7)	443 (100%)

なお、The Century of Prose Corpus (COPC) の 1710-1779 年間のテキストの調査では、*the reading the book* はすでに大幅な減少傾向が見られる。従って、両構文の並立する期間は比較的短いものと思われ、その期間の下限は Visser の言う “the end of the nineteenth

century” よりもっと早い時期ではないかと推測される。これは PPCEME と COPC では解決できない問題である。

いずれにしても、論文執筆時に PPCEME を構成するテキストジャンルに注意を払わなかったことは、通時的英語研究にコーパスを利用する者としては大いに反省させられることであった。