NPCMJ を用いた文構造の出現頻度に関する調査：主語省略文と受身文を例に

理論・対照研究領域　プラシャント・パルデシ　長崎 郁

大学共同利用機関法人 人間文化研究機構 国立国語研究所 National Institute for Japanese Language and Linguistics NINJAL

NPCMJ とは

国立国語研究所共同研究プロジェクト『統語・意味解析コーパスの開発と言語研究』では、2016 年度より日本語の統語解析情報付きコーパス **NPCMJ**（**NINJAL Parsed Corpus of Modern Japanese**）の構築を進めている。このコーパスは、現代日本語の書き言葉と話し言葉のテクストに対し文の統語・意味解析情報を付与し、多様な日本語の機能語や句構造、節の諸類型および複雑な構文を大量の言語データから検索・抽出して研究に活用することを目的としている。2021 年 3 月現在、約 6 万 7000 文（6 万 7000 ツリー）が公開され、2022 年 3 月末にさらに 1 万文が追加される予定である。

主語省略文と受身文の出現頻度の調査*

世界の諸言語、特に英語と対照した場合の日本語のコミュニケーション上の特徴とされてきた文法現象の中から主語省略文と受身文をとりあげ、NPCMJ を利用して量的な観点から分析し、その使用実態を明らかにした。

主語省略文：話しことばと書きことばという区別から見ると、100 文あたりの主語省略文の頻度は話しことば（日常会話と国会会議録）の方が、書きことば（新聞記事、エッセイ、フィクション、法律文）よりも高い。ただし、同じ話しことばであっても、日常会話と国会会議録では頻度に差がある。このような差を生む要因は、国会での発言には原稿を読みあげていると考えられるものも多く、その点で国会会議録のデータはより書きことばに近い側面をもつことにあると考えられる。書きことばの中では新聞記事の頻度が最も高いが、新聞記事における主語省略文には、見出し文が一定数含まれており、このような用例をのぞけば、頻度は幾分低くなるであろう。

表 1：主語省略文のジャンル別出現数と頻度

	出現数	100文あたり
日常会話	610	83.7
国会会議録	877	67.7
新聞記事	2015	42.6
エッセイ	1178	35.5
フィクション	1794	34.2
法律文	45	33.3
翻訳	2036	28.7

受身文：まず調査対象データ全体の中で 3 タイプの受身文（直接受身文、持ち主の受身文、間接受身文）の出現数に差があるかを調べた。その結果、直接受身文の出現数がほかの 2 つのタイプよりもかなり多く、受身文全体の 95%以上を占めていることが明らかになった。

表 2：直接受身文のジャンル別出現数と頻度

	出現数	動詞述語節 100節あたり	1,000語あたり
法律文	62	8.5	8.9
エッセイ	640	6.6	7.2
新聞記事	615	5.9	7.1
国会会議録	183	5.6	5.6
翻訳	659	4.4	5.5
フィクション	399	2.9	3.6
日常会話	4	0.3	0.3

直接受身文の使用において注目されるのは、日常会話における頻度の低さである。本調査のみで一般化することは難しいものの、ほかのジャンルの頻度は日常会話のおよそ 10〜30 倍となっている。話しことばであっても国会会議録のデータは書きことばに近い側面をもつことを述べたが、このことは、国会会議録における直接受身文の頻度が日常会話よりもかなり高く、書きことばのジャンルである新聞記事やエッセイと同程度であることにも反映されていると考えられる。書きことばの中で直接受身文の頻度が最も高いのは法律文であるが、法律文における直接受身文は、名詞修飾節で用いられることが多く、およそ 9 割を占めている。このような受身文の使用は、主語省略文と並んで、法律文の文体を特徴づける要素のひとつとなっている。

*調査の詳細は、2022 年 3 月刊行予定の窪薗晴夫・朝日祥之（編）『言語コミュニケーションの多様性』（くろしお出版）に収録

謝辞　本発表は国立国語研究所共同研究プロジェクト「統語・意味解析コーパスの開発と言語研究」および「対照言語学の観点から見た日本語の音声と文法」による研究成果の一部を報告したものである。

統語・意味解析コーパスの開発と言語研究
プロジェクトサイト
http://npcmj.ninjal.ac.jp

YouTube 動画
NPCMJ: 概要と検索方法
https://www.youtube.com/playlist?list=PLZfZgVvFbh1ZLsndcOVYaS-z3GFkH5Any