定例研究会要旨

日時：平成 21（2009）年 12 月 2 日　18:00ー20:00
会場：東京外国語大学　語学研究所
題目：「コーパスを利用した近代語研究～太陽コーパスと近代文語 UniDic～」
発表者：小木曽 智信（国立国語研究所 言語資源研究系准教授/日本語学）

国立国語研究所で作られた近代日本語のコーパスである『太陽コーパス』と，近代文語文の形態素解析を可能にする「近代文語 UniDic」の概要とその利用例を紹介した．

『太陽コーパス』は明治・大正期の総合雑誌『太陽』のうち，創刊の 1895 年と 1901 年から 8 年おきに 1909 年，1917 年，1925 年の 5 年分の約 3400 記事，計約 1445 万字のテキストを収録している．雑誌『太陽』は当時最もよく読まれた雑誌であり，多くの著名人を含む多彩な執筆者により，自然科学や社会科学から文学作品まで，広汎な分野にわたる記事が掲載されている．結果として，ある程度ジャンルのバランスのとれたデータとなっており，当時の言語の実態を反映した資料となっている．

『太陽』の本文は今日のテキストとは異なり，不統一な句読法，臨時的熟字訓を含む多様なふりがな，仮名遣いの不統一，濁点付与の不徹底，難解な漢字や異体字の使用，数多い誤植など，特に表記上多くの問題がある．このような原文の情報をできる限り残しつつ，検索に適したデータとしてとりまとめるために，太陽コーパスでは XML を利用して独自のタグセットによるタグ付けを行っている．表記面の情報のほかに，記事の著者やジャンル，引用元などの言語研究に必要な情報が付加されている．この XML ファイルを活用するため，各種 XSLT スタイルシートを同梱しており，これを用いることで原文から必要な情報を抜き出したり，ファイル形式を変換して利用することが可能である．

太陽コーパスの一般的な利用法として，付属の全文検索システムひまわりを用いた用例検索がある．その例として，(1)「拉致」の用例収集，(2)「婦人」と「女性」の用例数の推移，(3)「しょうがい（障害）」の表記の推移，の三つのケースを紹介した．

(1)「拉致」の用例では，文脈から見て初期の例は「引っ張ってくる」「連れてくる」といった意味であり，強制的な意味合いを含んでいないと考えられることを示した．たとえば，「即ち大に門戸を開いて，眞個政治家の資質ある人士を拉致するの外に無いのである」（1909 年 5 号「政党の革新」建部遯吾）などはその例である．

(2)「婦人」と「女性」の用例数の推移では，もともとほとんど用いられていなかっ

た「女性」の語が 1925 年から急激に増えていることを示した．今日，「婦人」の語が廃れあまり用いられなくなっているが，この変化がこの頃始まること，「婦人」と「女性」をあわせた用例数も激増しており，大正期の女性の社会進出が背景にありそうなことを指摘した（図 1）．

(3)「しょうがい（障害）」の表記の推移では，マスコミ等で見られる“「障害」は，戦前は「障碍」と書いていたが，当用漢字の制限により「碍」の字が用いられなくなったために新たに「障害」と書かれるようになった”とする説が誤りであり，太陽コーパスでも「障害」が最も多く用いられていることを示した（図 2）．

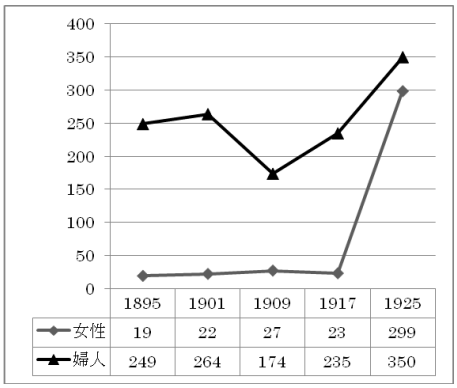

図 1

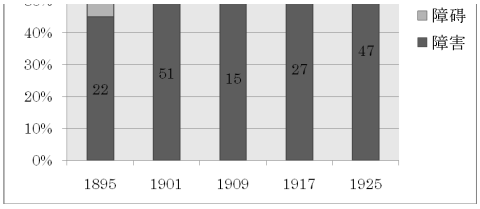

図 2

以上のように文字列検索でもさまざまな利用が可能であるが，表記上区別がつきにくい語や不特定の語の検索・集計ができないなど，文字列検索には限界がある．コーパス全体の延べ語数・異なり語数も不明であり，検索可能な語であってもコーパスの語彙全体の中に占める割合などを求めることができない．しかし，形態素解析を用いることで，こうした問題を解決することができる．

太陽コーパスに形態素解析を施すことを目的に開発を行ってきた近代文語 UniDic は，広く用いられている形態素解析器 ChaSen，MeCab で利用可能な形態素解析辞書である．現代語用 UniDic の見出し語に加え，活用語の文語形や旧字旧仮名の表記などを追加し，太陽コーパスの一部のデータなど近代文語文のコーパスを機械学習に用いている．これにより，おおよそ 95〜98 パーセントの精度で形態素解析を行うことが可能になった．もとより完全な解析結果ではないが，テーマを選んだり，結果に人手修正を加えたりすることにより研究に利用することが可能になった．発表では形容動詞の連体修飾の形（例：「重大の」「重大なる」「重大な」）を例に，形態素解析を用いた集計結果を示し，形態素解析を用いた研究の可能性について論じた．