『現代日本語書き言葉均衡コーパス』新聞サブコーパスに対する新聞記事情報の付与

加藤祥，森山奈々美，浅原正幸

『現代日本語書き言葉均衡コーパス』（以降 BCCWJ）の新聞サブコーパス（以降 PN）に含まれる新聞データに，新聞記事情報を付与した。具体的には，BCCWJ の PN ファイル全サンプル 1,473 サンプル（非コアも含む）について，BCCWJ 構築時にサンプリング対象となった実際の新聞紙面を手作業で確認し，サンプル内に含まれる記事単位の分割を行うとともに，記事ごとの掲載紙面（記事分類情報）や記事種別情報を取得した。

新たに付与した情報は，(i)サンプル内の各記事の開始位置，(ii)サンプル内の記事番号と各記事の完結・未完結，(iii)記事の掲載紙面名，(iv)記事種別：1（評論・催し案内・人・色などの別）・2（一般報道・連載・解説などの別），(v)国内・海外・国際の別，(vi)記事内容：大分類（政治・文化・スポーツなどの別）・小分類（大分類がスポーツであれば，陸上・野球・武道などの別）である。

本発表では，新聞記事情報付与の設計と実際の付与作業，作業結果の基礎統計データについて報告する。記事の単位で紙面情報や記事内容・記事種類などの情報を付与することにより，PN に含まれる記事別の語彙調査や文体分析などの詳細な検索や分析が可能となった。付与された記事分類と記事種別の情報を用いることにより，たとえば，国内の事件報道のみの抽出や，料理レシピの抽出，連載小説の除外などが容易である。

また，BCCWJ には，意味情報や比喩情報など様々な情報が付与されているため，本作業結果を重ね合わせた集計も可能となった。分類語彙表番号付与データと記事内容を重ね合わせ，新聞の記事内容別に意味分野の分布を集計した例も紹介する。