定例研究会要旨

日時：平成 24（2012）年 7 月 11 日　17:40～19:40
会場：東京外国語大学　語学研究所
題目：「通時コーパス構築上の諸問題」
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/コーパス日本語学）

国立国語研究所では，共同研究プロジェクト「通時コーパスの設計」（プロジェクトリーダー：近藤泰弘）の下で，日本語の通時コーパスの設計と実装に関する研究を行っている．本発表ではこの通時コーパスを構築するうえでの問題について論じた．

通時コーパスを構築する作業の流れはおおよそ次のようになる．

1. 本文選定
2. 電子テキスト化
3. 外字の処理
4. 文書構造のタグ付け
5. テキストの解析前処理
6. 形態素解析
7. データベースへのインポート
8. 解析誤り修正
9. （長単位・文節解析）
10. （係り受け解析）
11. XML 出力　／　アプリケーションでの利用

以下，この流れに沿って課題を見ていくことにする．

1.の本文選定にあたっては，日本語学の研究資料として価値の高いものを選ぶことは当然であるが，それだけでなく，定評のある校訂済み本文であることも要求される．大部分の資料は，その成立時期から大きく下った時代の写本によって伝えられたものであるため，本文批判を経て作られた精確な本文である必要がある．また，後述する形態素解析などの処理を行うためには，読みやすい漢字仮名交じりに直された本文であることが望ましい．

2.の電子テキスト化では，テキスト入力の作業そのものは外注業者に任せることも可能であるが，内容の読解が容易でなく一般的でない文字も多いため，その後の校正作業を専門知識を持つ者が行う必要がある．単に文字を入力するというだけでなく，規格として標準化された符号化文字集合に準拠して電子化する必要があるため漢字と文字コードに関する知識が必要とされる．

そして，一般的なパソコンでは扱いが難しい文字が多く出現するため 3.の外字処理が必要とされる．通時コーパスでは，「現代日本語書き言葉均衡コーパス」と同様，JIS X0213

を利用しており，これによりかなりの程度の漢字をカバーできる．しかし，たとえば『今昔物語集』では，のべ約 1000 字，異なり約 200 字程度が表現できない．コーパスでは利用の便を考えると「=」を用いた処理は避けたいため，規格外字は極力規格内の別字で代用する方針ですすめている．

4.の文書構造のタグ付けとは，タイトル・記事・注などの文書構造に関する情報を残したり，段落・引用・文などの言語構造に関わる情報を付与する作業である．マークアップ言語 XML を用いて各種の情報をタグ付けしている．タグ付けは，要らない情報をそぎ落とすとともに必要な情報を付与する作業であるから，通時コーパスにとって必要な情報とは何かを考えて設計を行う必要がある．TEI などの国際的な標準も視野に入れながら，通時コーパスに必要な情報を吟味し，文書構造の設計を行っている．

5.のテキストの解析前処理は，この後に行う形態素解析のために，処理がしやすい（そして人間にも読みやすい）形にテキストを整える作業である．通時コーパスのテキストでは，返り点が含まれていたり，濁点が付けられていなかったり，踊り字が使われていたりする．このままでは処理が難しいため，通常のテキストの形に整える作業が必要となる．濁点の自動付与などの研究も行っているが，いずれも最後には人手による確認と修正が必要である．

6.の形態素解析は，テキストを単語に区切り，読み・代表表記・品詞・活用などの情報を付与する作業である．通時コーパスでは，形態素解析器「MeCab」と形態素解析辞書「近代文語 UniDic」「中古和文 UniDic」などを用いて処理を行っている．典型的なテキストであれば自動処理で 96%程度の精度で解析を行うことが可能になっているが，これらの辞書ではうまく解析でいない資料も少なくないため，新たな辞書の整備や解析時の工夫が必要となっている．

形態素解析済みのデータは，7.のデータベースへのインポートによって形態論情報データベースに格納したのち，人手により専用のツールを使って 8.の解析誤り修正を行う．解析結果の修正は原文を読解したうえで，文法的な知識に基づいて行う必要がある高度な作業である．

この後に，9.（長単位・文節解析），10.（係り受け解析）による高度な情報の付与を行う個とを検討しているが，現在は研究段階である．

修正が終わったデータは，11. の XML 出力／アプリケーションでの利用という形で，最終的な配布用データや，利用しやすいサービスとして外部に提供することになる．これまでに，コーパスを利用するアプリケーションとして通時コーパス版の「中納言」を開発し，共同研究者向けのサービスを開始している（発表会場でデモを行った）．このほか，コーパス管理ツール「茶器」の紹介を行った．