言語資料の共有、利用を支援する環境構築に関する研究

代　表　者：山口昌也
（独立行政法人国立国語研究所研究員）

研究共同者：小木曽智信
（独立行政法人国立国語研究所研究員）

間淵洋子
（独立行政法人国立国語研究所特別奨励研究員）

研究成果要約

研究活動概要

本研究の目的は、電子化された言語資料の共有、利用を支援するツール（ソフトウェア群）を構築することである。本研究では、コンピュータに関する専門的な知識を持たない言語研究者のことを考慮し、（1）理解することが容易な言語資料の記述形式（共有記述形式）の提案、（2）提案した形式で記述した言語資料を、検索システムで利用できる形式に変換するツールの開発、（3）大量の言語資料を処理することができる全文検索システムの開発を行った。これらの研究により、多くの研究者が共有可能な言語資料を容易に作成するための基礎的な環境を実現することができた。また、本研究の成果物である全文検索システム、既存の言語資料を共有記述形式に変換するための変換ツール、さらに、共有記述形式に変換した言語資料（著作権上問題ない資料）を一般に公開することができた。

成果概要

（1）言語資料を共有するための記述形式として、次の二つの形式を定義した。これらは、XML の DTD（Document Type Definition）として定義した。

- 共有記述形式（書き言葉用）
- 共有記述形式（話し言葉用）

（2）言語資料を共有記述形式に変換するため、次の二つのツールを開発した。

- 変換ツール『えだまめ』（変換の簡便性を追及した変換ツール）
- 『簡易変換ツール』（柔軟性の高い文字列変換機能を持つ変換ツール）

（3）次の既存の言語資料を、共有記述形式に変換した。

- 書き言葉の言語資料（『青空文庫』『日本古典文学本文データベース』（国文学研究資料館）など2635作品）
- 話し言葉の言語資料（『BTSJ による多言語話し言葉コーパス』（宇佐美 2003）など 4 談話資料）

（4）全文検索システム『ひまわり』サーバ版、クライアント版を構築した。

成果活用について

上記の成果物のうち、変換ツール『えだまめ』、全文検索システム『ひまわり』、共有記述形式の言語資料（著作権上問題のないもの）については、すでに国語研究所の Web

ページで公開している。共有記述形式の DTD、および、著作権上問題のある言語資料については、（変換方法を公開するなどの方法を用いて）順次 Web ページ上に公開する予定である。

今後の研究課題

今後の研究課題として、次の 3 点を挙げる。

- **共有記述形式の記述能力の追加的評価：** 既存の言語資料を変換することにより、部分的にではあるが、評価を行った。今後は、実際の研究者が共有記述形式で言語資料を作成できるかを確認したいと考える。
- **変換ツールの能力の向上：** 変換ツール『えだまめ』は、マウス操作だけでプレーンテキストから共有記述形式への変換を行うことができる。しかし、共有記述形式のすべてのタグを生成できるわけではない。そこで、プレーンテキストからより多くのタグを容易に生成する仕組みを考える必要がある。
- **サーバ版『ひまわり』の能力向上：** 基本的な全文検索を実現することができたが、大量の検索結果が得られた場合、クライアントに対するレスポンスが悪くなるという問題がある。今後、すべての検索結果をクライアントに送信するのではなく、部分的に送信する機能を実装するなどの対策が必要である。