様式 7-3
助成番号： 05-B-0033
2007 年 04 月 30 日

言語資料の共有，利用を支援する環境構築に関する研究

山口昌也，小木曽智信，間淵洋子
独立行政法人 国立国語研究所

1　はじめに

本研究の目的は，電子化された言語資料の共有，利用を支援するためのツールを実現することである。本研究では，利用者として，日本語に関する言語研究者，特に，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を想定する。

本研究の背景として，実例を基にした言語研究における言語資料の作成の問題がある。実例を基にした言語研究を行う場合，研究対象となる言語資料を準備することは必要不可欠である。しかし，個人の研究者が言語資料を作成する場合，独自の形式で作成することになりがちである。この状況には，次の二つの問題があると考える。

- 作成した言語資料が個人の研究内にとどまり，他の研究者と共有することができないこと
- 作成した言語資料を利用する手段（例えば，検索ツール）を研究者個人が作成しなければならないこと

そこで，本研究では，言語研究者が電子化された言語資料の共有，利用を支援するため，次の三つの事柄を実現することにする。

サブテーマ 1:　言語研究者が共有可能な言語資料の形式の設計
サブテーマ 2:　言語研究者が共有記述形式の言語資料を容易に作成できるようにするための支援ツールの実現
サブテーマ 3:　共有化された言語資料を効率的に利用するツールの実現

まず，サブテーマ 1 では，言語研究にとって必要不可欠な情報を選りすぐり，研究者が理解しやすく，かつ，容易に記述可能な言語資料の形式 (共有記述形式) を設計する。設計にあたっては，既存の記述形式，および，言語資料について調査を行い，言語資料に付与する情報として，言語研究者が何を必要としているかを明らかにする。本研究ではこの調査に基づき，書き言葉用，話し言葉用の 2 種類の共有記述形式を定義する。

次に，サブテーマ 2 では，共有記述形式の言語資料の作成支援ツールとして，二つのツールを作成する。一つは，操作の簡便性に焦点を当てた変換ツール『えだまめ』である。このツールでは，誰でも共有記述形式の言語資料が作成できることを目指す。もう一つは，『えだまめ』よりも操作の難易度は高いが，より柔軟な文字列の変換ができるように設計された『簡易変換ツール』である。このツールは，正規表現による変換規則集合に基づき，既存の言語資料を共有記述形式に変換する。

サブテーマ 3 では，共有化された言語資料を効率的に利用するためのツールとして，サーバ・クライアント型の全文検索システムを実現する。このシステムは，複数のコーパスを集中管理し，ネットワークを介した全文検索を行うことができる。システムの実現は，筆者らが開発した全文検索システム『ひまわり』を拡張することにより行う。

本報告書の構成は，次のようになっている。まず，2 節で，本研究の流れとそれに伴う研究成果を概説する。3 節では，サブテーマ 1 として，言語資料を共有化するための共有記述形式の設計と，共有記述形式による既存の言語資料の記述を行う。4 節では，サブテーマ 2 として，共有記述形式の言語資料への変換ツールの説明を行う。5 節では，サブテーマ 3 として，サーバ・クライアント型全文検索システムの説明を行う，そして，6 節で本研究全体の評価を行い，7 節でまとめを述べる。

2　研究の流れと成果物の概要

本研究の流れを図 1 に示す。この図では，サブテーマごとの関係とその成果物を併せて示している。赤の☆が付与されている成果物は，国語研究所の Web ページ[*1]で公開済みのもの (部分的に公開しているものも含む)，緑の☆の場合は，今後 Web 上に公開を予定している成果物である。

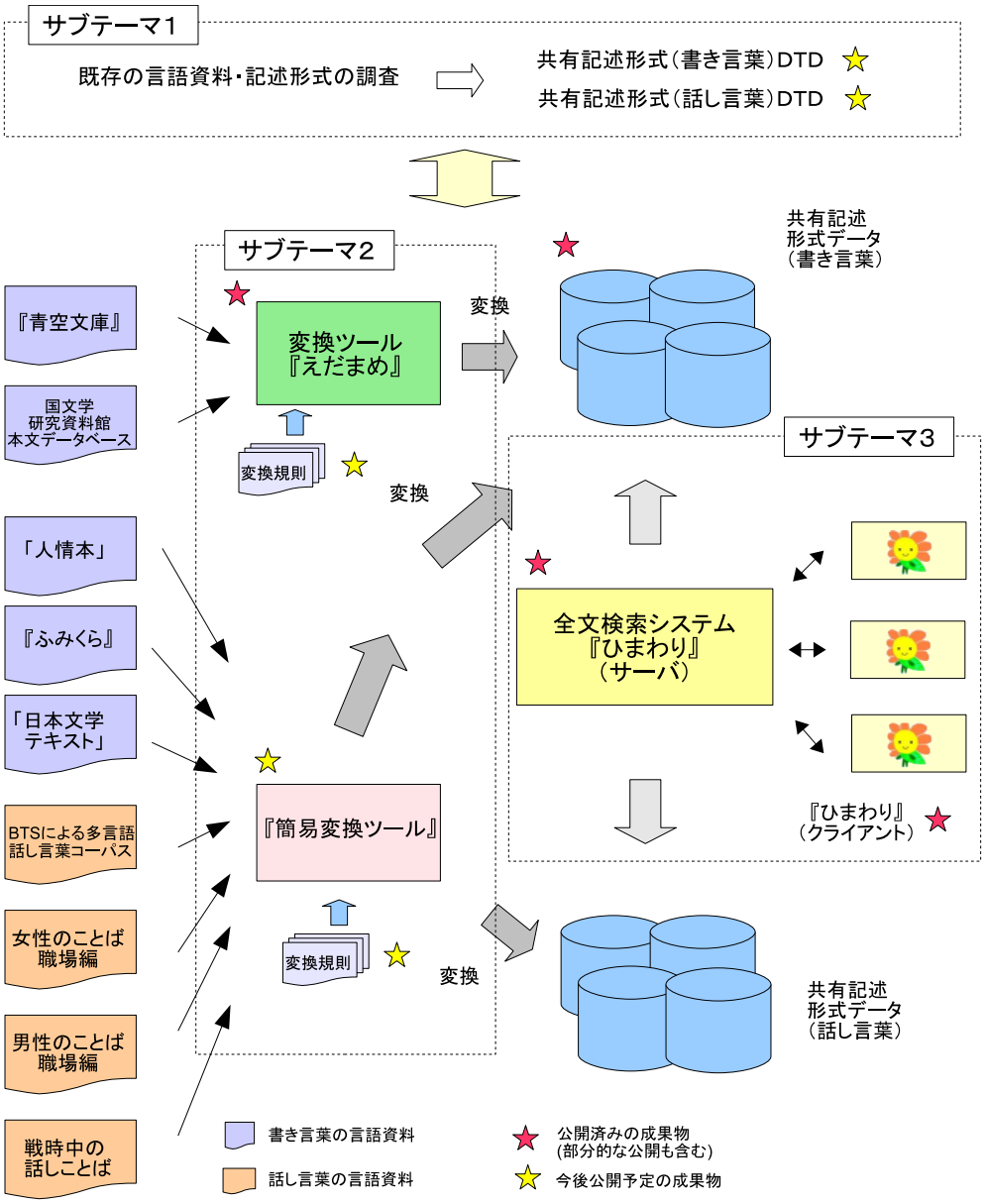

図 1　研究の流れ

[*1] http://www.kokken.go.jp/lrc

■サブテーマ１： サブテーマ１では，既存の言語資料・記述形式の調査を行い，調査結果に基づいて，書き言葉，話し言葉用の二つの共有記述形式を定義した。どちらの記述形式の定義も XML の DTD (Document Type Definition) により規定した。サブテーマ１における成果物は，次のとおりである。なお，サブテーマ１は，間淵，山口，小木曽が共同で行った。

- 共用記述形式（書き言葉用）の仕様書，および，DTD ［今後，公開予定］
- 共用記述形式（話し言葉用）の仕様書，および，DTD ［今後，公開予定］

■サブテーマ２： サブテーマ２では，『えだまめ』『簡易変換ツール』の二つのツールを作成した。この二つのツールを利用し，既存の言語資料 (図 1 左) を共有記述形式に変換する。さらに，この変換によって，サブテーマ１で定義した共有記述形式の評価を行い，不備があれば，仕様にフィードバックする。また，この過程で変換ツール側の改良も実施した。

言語資料の変換は，２通りの方法で行う。まず，言語資料のうち，XML で記述されているものについては，個々の資料ごと XSL スタイルシート (変換規則) を記述し，変換ツール『えだまめ』で共有記述形式 (XML) に変換する。一方，独自の形式で記述されている言語資料については，『簡易変換ツール』用の変換規則を個々の言語資料ごとに作成し，共有記述形式に変換する。サブテーマ２は，小木曽，山口が共同で行った。成果物は，次のとおりである。

- 変換ツール『えだまめ』 (国語研究所の Web サーバで公開済み)
- 『簡易変換ツール』 ［今後，公開予定］
- 変換結果の共有記述形式の言語資料 (詳細は，3.4 節を参照) ［国語研究所の Web サーバで公開済み］
  - 『ふみくら』
  - 「日本文学テキスト」
  - 「人情本」
- 共有記述形式への変換規則 ［今後，公開予定］
  - 『えだまめ』用 (『青空文庫』「日本古典文学本文データベース」)
  - 『簡易変換ツール』用 (『BTS による多言語話し言葉コーパス』『女性のことば・職場編』『男性のことば・職場編』『戦時中の話しことば』)

■サブテーマ３： サブテーマ３では，既存の全文検索システム『ひまわり』を複数・大規模なコーパスへ対応できるよう拡張する。さらに，『ひまわり』にサーバ機能，および，クライアント機能を追加する。さらに，『ひまわり』および共有記述形式の評価として，サブテーマ２で変換した言語資料を『ひまわり』に搭載し，動作を検証する。

成果物は，次のとおりである。「一部公開済み」となっているのは，複数・大規模なコーパスへの対応を行った『ひまわり』ver.1.3 β を公開しているためである。なお，サブテーマ３は，山口が行った。

- 全文検索システム『ひまわり』（サーバ版） ［一部公開済み］
- 全文検索システム『ひまわり』（クライアント版） ［一部公開済み］

3 共有可能な言語資料形式

3.1 設計方針

本研究では，共有記述形式の設計方針として，次の二つのことを掲げる。

(1) 研究者が理解しやすく，かつ，容易に記述可能なこと
(2) 計算機システムにとって扱いやすい形式であること

一つ目の設計方針は，共有記述形式の設計において，最も重視することである。なぜならば，共有記述形式は，個人の研究者が自分で言語資料を作成する際に利用することを想定しており，記述形式が理解しやすく，資料を記述しやすいことが重要になるからである。さらに，言語資料を共有するということは，他の研究者が作成した言語資料を理解した上で利用することを意味する。したがって，理解のしやすさという要素は非常に重要である。

以上のことから，可能な限りシンプルな記述形式とすることを目指す。ただし，研究者が言語資料に対して必要とする情報は，研究者ごとに多種多様であり，個々の研究目的に適した記述を行えることが記述形式には求められる。そこで，シンプルであり，かつ，拡張性を有した記述形式となるよう留意する。

二つ目の設計方針は，共有記述形式の言語資料を，さまざまな計算機処理が容易にできるようにすることを意図している。本研究では，すでに述べたように，共有記述形式の言語資料を全文検索システムで利用することを想定しているが，自動形態素解析を行い，語彙の分析をするなど，全文検索システム以外にもさまざまな利用方法が考えられる。

そこで，共有記述形式では，XML (Extensible Markup Language) を用いて言語資料を記述する。XML は，文書記述言語の国際的な標準規格であり，データ処理一般に広く利用されているほか，コーパスを作成する際にも標準的に用いられている。そのため，データの整合性のチェックや形式の変換など，言語資料を扱う上で基本的なアプリケーションを容易に入手することが可能であり，自分の目的にあったアプリケーションを構築しやすい[*2]。

3.2 既存の言語記述形式と言語資料

実際の設計を行う前に，既存の記述形式と言語資料を概観してみよう。

代表的な記述形式としては，TEI (Text Encoding Initiative)[3]，CES (Corpus Encoding Standard)[4]，KOKIN ルール [1, 2] が挙げられるだろう。まず，TEI は，コーパスのための汎用電子化フォーマットであるため，研究者の多様な要望に耐えうる記述能力を持っている。しかし，汎用ということもあり，仕様が複雑であり，仕様を理解し，実際に言語資料を記述するのが容易ではない。一方，CES は TEI よりもシンプルな仕様であるが，適用範囲として，言語工学やその応用を指向しており，言語学的分析を目的とした言語資料に CES をそのまま適合することは難しい。次に，KOKIN ルールだが，この記述形式は日本語の古典テキストを記述することを目的とし，SGML として規定されている。したがって，古典テキストに特化した仕様であること，また，現在では XML での記述が主流であることを勘案すると，共有という面からは利用しにくい。

次に，研究者がどのような言語資料を作成し，その言語資料にどのような言語学的な情報が付与されている

[*2] 実際，本研究を行うにあたって利用したソフトウェアの多くは，フリーソフトウェアである。

かを見てみよう。ここでは，実際に共有を意図した言語資料の例として，テキストデータ (日本語の言語資料となりうるものに限る) を公開している Web サイトを中心に[*3]調査した。調査した Web サイトの例を次に示す。

- 青空文庫 (http://www.aozora.gr.jp/)
- 「ふみくら」(http://www.fumikura.net/)
- J-TEXTS (http://www.j-texts.com/)
- Japanese Text Initiative (http://etext.lib.virginia.edu/japanese/index.euc.html)
- digital 西行庵 (http://www.saigyo.org/)
- 日本語の歴史と日本語研究の歴史 (http://www.ne.jp/asahi/nihongo/okajima/)
- 日本ペンクラブ　電子文芸館 (http://www.japanpen.or.jp/e-bungeikan/home.html)
- 謡曲三百五十番集入力 (http://www.kanazawa-bidai.ac.jp/ hangyo/utahi/)

調査の結果をまとめると次のようになる。

- 公開されている資料の種別
  - 著作権の問題もあり，著作権が切れたデータ，特に古典テキストを扱うことが大部分である。
  - ただし，談話資料も少数 (4 点，詳しくは 3.3.2 を参照) であるが見受けられる。
- 主な付与情報 ([] 内の数字は，調査した 57 Web サイトの中で付与していたサイト数)
  - 書誌情報 (作品名，著者，底本情報，入力者，校訂者など) ［全サイト］
  - 注記 (校注，入力者注，ママ注など) ［23 サイト］
  - 文字情報 (外字，読み取り不能など) ［33 サイト］
  - ルビ ［26 サイト］
  - 見出し［11 サイト］
  - 位置情報 (行番号，ページ番号など)［9 サイト］

3.3 共有記述形式

3.2 節での調査に基づき，共有記述形式の仕様は，既存の記述形式をそのまま利用するのではなく，標準的な記述形式との連係が可能となるように考慮しつつ，独自に設計していくことにする。また，本研究では，書き言葉と話し言葉 (談話資料) を記述対象の言語資料とし，二つの共有記述形式を個別に定義することにする。このうち，書き言葉の共有記述形式については，前節の調査結果に基づき，古典テキストを扱えるよう配慮する。

ここでは，まず，二つの共有記述形式で共通する仕様として，文字集合，文字符号化方式，文書記述言語について，次のように定める。個々の仕様の固有部分については，後続の節で順次説明する。

文字集合： Unicode
文字符号化方式： UTF-16 (little endian)
文書記述言語： XML

[*3] 書籍などに付随する CD-ROM なども含む。

3.3.1 共有記述形式 (書き言葉)

書き言葉の共有記述形式は，24 個の XML タグから構成される。主な XML タグを表 1 に示す。なお，紙面の関係上，各タグの属性についての説明は，省略する。また，詳細な仕様は，後日 Web ページ上に公開する予定である。

表 1　共有記述形式 (書き言葉) のタグ一覧 (一部)

タグ名	内容
コーパス	一つのコーパスを表す。
記事	同一著者による，同一テーマの一まとまりの文書。凡例や著作件表示など，本文に付随する文書要素も含む。
テキスト	「記事」要素内の本文 (ヘッダとフッタ以外) を表す。
ヘッダ	凡例，著作権表示など，本文より前に出現する文書要素を表す。
フッタ	謝辞，著作権表示など，本文より後に出現する文書要素を表す。
頁	頁・丁の区切り位置を表す。
注	入力者，校訂者，作者による本文に対する注記 (誤字，脱字，衍字，訂正など) を表す。
引用	和歌，俳句，歌謡，会話などの引用を表す。
外字	規定された文字集合に含まれない外字を表す。
ルビ	ルビを表す (タグの属性で，左ルビを記述することも可能)。
ブロック	「テキスト」要素内の文書要素の種類を規定する汎用ブロックタグ (論理行以上の文書要素用)
span	「テキスト」要素内の文書要素の種類を規定する汎用 inline タグ (行中に現れる文書要素用)

この仕様の特徴の一つは，設計方針のとおり，可能な限りタグの数を減らし，理解・記述しやすくしていることである。ただし，3.2 節での調査に基づき，「注」「外字」「ルビ」など多くの研究者が必要としているタグについては，仕様の中に含めた。また，この記述形式では定義されていない文書要素を記述する要求に答えるため，「ブロック」「span」という汎用のタグを用意した。「ブロック」タグは，1 行以上の論理行からなる文書要素を表す場合に使用し，「span」タグは，行中に含まれる文書要素を表す場合に使用する。例えば，本記述形式には，図のキャプションを表すための専用のタグが用意されていないが，「ブロック」タグを下 (左) のように使用することにより，「キャプション」を表現する。また，下 (右) の「span」タグは，原資料で強調表示されている文書要素を表現している例である。

<ブロック 種別="キャプション">
：（キャプションをここに記述する）
</ブロック>

<span 種別="強調">先夜の捕物は

もう一つの特徴は，本文を表す「テキスト」とそれ以外の要素「ヘッダ」「フッタ」とを明確に区別している点である。これにより，本文とそれ以外の要素 (例：凡例，著作権表示，謝辞など) を明示的に区別することができる。したがって，本文に影響を与えることなく，著作権表示を追加したり，本文だけの検索，言語資料全体の検索といったことが可能になる[*4]。

[*4] 例えば，本研究の成果物として一般公開されている『ひまわり』用の言語資料は，筆者らの著作権表示を追加するとともに，「テキスト」要素内だけを検索するように設定されている。

図 2 に，共有記述形式を利用した記述例を示す。この例は，川上眉山の「ゆふだすき」[*5]を共有記述形式に変換した結果である。

```
<コーパス 名前="日本語文学テキスト">
<記事 タイトル="ゆふだすき" 著者="川上眉山" 底本="＊">
<ヘッダ>
Copyright (c) 2007  Masaya YAMAGUCHI, Sae UENO, Miwa FUJIMOTO
This document was transformed to HIMAWARI format from the original document.
</ヘッダ>
<テキスト 種別="本文">
ゆふだすき　川上眉山<行 /><行 />
(一) <行 /><行 />
いや、驚いたよ君、何ものほほんで歩いて居た<行 />
譯ぢゃなかッたが、不意に横ッ手から、<行 />
<引用 種別="会話" 引用元="">「あら、まァ、梅原さんぢゃァありませんの。」</引用><行 />
と甲の高い、調子の走ッた、<r rt="けしゃう">化生</r>の者の叫び聲<行 />だ。
                :
                :
</テキスト>
<フッタ>
岡島昭浩入力
底本『現代日本文学全集 第七編 柳浪・眉山・緑雨集』改造社 昭和 4.3.1
ただし、底本ではほぼ総ルビ。
日本文学等テキストファイルへ
                :
</フッタ>
</記事></コーパス>
```

図 2　共有形式（書き言葉）による記述例

3.3.2　共有記述形式 (話し言葉)

話し言葉の共有記述形式は，15 個の XML タグから構成される。主な XML タグを表 2 に示す。なお，紙面の関係上，各タグの属性についての説明は，省略する。また，詳細な仕様は，後日 Web ページ上に公開する予定である。

共有記述形式の定義は，次の四つの既存の談話資料を統一的に記述できるようにするという観点から行った。資料 (1) は，日本語話者同士の会話，日本語話者と日本語学習者との会話を書き起こした資料 (約 28 時間

[*5] 岡島昭浩氏（大阪大学大学院文学研究科国文学・東洋文学講座）が http://www.let.osaka-u.ac.jp/ okajima/bungaku.htm）で公開しているデータを利用

表 2　共有記述形式 (話し言葉) のタグ一覧 (一部)

タグ名	内容
コーパス	一つのコーパスを表す。
記事	一連のひとまとまりの談話を表す。凡例や著作件表示など，談話に付随する文書要素も含む。
テキスト	「記事」要素内の談話部分 (ヘッダとフッタ以外) を表す。
ヘッダ	凡例，著作権表示など，談話部分より前に出現する文書要素を表す。
フッタ	謝辞，著作権表示など，談話部分より後に出現する文書要素を表す。
発話文	一つの発話を表す。
音声的情報	イントネーション，言い淀みなどなどの音声的情報 [7] を表す。
周辺言語情報	あいづちや笑いなど周辺言語情報 [7] を表す。
注	入力者，校訂者，作者による本文に対する注記 (誤字，脱字，衍字，訂正など) を表す。
引用	和歌，俳句，歌謡，会話などの引用を表す。
外字	規定された文字集合に含まれない外字を表す。
span	「テキスト」要素内の文書要素の種類を規定する汎用 inline タグ (行中に現れる文書要素用)

分) である。資料 (2)(3) は職場における話し言葉，それぞれ約 9 時間，約 21 時間分である。資料 (4) は，ラジオドラマの 78 冊分の台本である。

(1)『BTS による多言語話し言葉コーパス』[5, 6]
(2)『女性のことば・職場編』[8] の付属資料
(3)『男性のことば・職場編』[9] の付属資料
(4)『戦時中の話しことば　ラジオドラマの台本から』[10] の付属資料

図 3 に共有記述形式による記述例を示す。

話し言葉の共有記述形式では，「テキスト」要素中の構造は簡単であり，図 3 の例のように，テキストの中に発話文が繰り返される形式になる。

それに対して，談話資料における発話条件の多様性に対応して，「発話文」タグに多くの属性を持たせている。ただし，無秩序に属性を増やすと仕様が複雑になるため，談話という面から，可能な限り一般性のある情報を記述形式の仕様に盛り込むことにした。また，資料固有の情報については，汎用的な属性を用意して，そこに記述するか，仕様中の類似する属性に記述する。例えば，資料 (4) はラジオの台本なので，他の三つの資料にはない「台本番号」や「放送日」という情報を含んでいる。このうち，「台本番号」は「会話文」タグの「タイトル」属性に記述し，「放送日」については，汎用の属性である「備考」属性に記述した。その一方で，「発話文」タグにおける「話者名」「性別」「年齢」「相手名」といった属性は，談話資料という面からは一般性を持つので，仕様の中に取り込んである。

```
<記事 タイトル="台本番号:1, タイトル:三四郎（一）" 備考="放送日:1936.08.31">
<ヘッダ>
                :
</ヘッダ>
<テキスト>
<発話文 番号1="1" 番号2="1" 話者名="解説" 性別="" 年齢="" 相手名=""
        相手性別="" 相手年齢="" 場所場面="" 備考="">
これは明治四十年頃の物語であります。
</発話文>
<発話文 番号1="2" 番号2="2" 話者名="解説" 性別="" 年齢="" 相手名=""
        相手性別="" 相手年齢="" 場所場面="" 備考="">
その頃、大学は九月に新学期が始まりました。
</発話文>
                :
                :
</テキスト>
<フッタ>
                :
</フッタ>
</記事></コーパス>
```

図 3　共有形式（話し言葉）による資料 (4) の記述例

3.4　既存言語資料の記述

定義した二つの共有記述形式の記述能力を評価するために，実際の言語資料を共有記述形式で記述することを試みた。記述対象とした言語資料は，次のとおりである。

- 書き言葉の言語資料 (合計で，2635 作品)

  『ふみくら』　高木 元氏 (千葉大学文学部) が Web サイト「ふみくら」(http://www.fumikura.net/, 日本の十九世紀小説に関する研究資料) で一般に公開している言語資料 (『南総里見八犬伝』など１７点) [Web 上で一般公開]

  「日本文学テキスト」　岡島昭浩氏 (大阪大学) が Web サイト「日本語の歴史と日本語研究の歴史」(http://www.let.osaka-u.ac.jp/ okajima/bungaku.htm) で一般に公開している言語資料 (『更科日記』など 5 点)

  「人情本」　岡部嘉幸氏 (千葉大学文学部) が作成した言語資料で，人情本刊行会の資料を翻刻した電子データ (『仮名文章娘節用』など 6 作品)

  『青空文庫』　『青空文庫』(http::/www.aozora.gr.jp) で公開されているデータのうち，XHTML 化さ

れているデータ。今回は，[11] に付属する CD-ROM 中のデータ 2560 作品を利用した。

「日本古典文学本文データベース」 国文学資料館が Web 上で公開している「日本古典文学本文データベース」に収録されている 47 作品

- 話し言葉の言語資料 (3.3.2 節で示した四つの談話資料)

以上の言語資料は，4 節で述べる変換ツール『えだまめ』と『簡易変換ツール』を用いて，変換した。上記の資料のうち，XML で記述されている『青空文庫』「日本古典文学本文データベース」のデータは，XSL スタイルシートを記述して，『えだまめ』で変換した。一方，その他の言語資料は，独自の形式で記述されているので，個々の資料ごとに変換規則を作成し，『簡易変換ツール』で共有記述形式への変換を行った。

最後に，変換された言語資料を 5 節で述べる全文検索システム『ひまわり』で検索できるか確認した。

なお，著作権上，再配布しても問題のない，『ふみくら』「日本文学テキスト」「人情本」は，変換後の共有記述形式化されたデータを一般に公開した。また，残りの『青空文庫』「日本古典文学本文データベース」，四つの談話資料に関しては，著作権上再配布することができないので，変換方法を Web ページで公開する予定である。

4 言語資料作成支援ツールの実現

4.1 『えだまめ』の設計と実現

4.1.1 『えだまめ』の概要

『えだまめ』は，既存の言語資料を全文検索『ひまわり』用のデータに変換するためのツールである。『えだまめ』の設計方針は，「可能な限り容易」に変換を実現することである。「限りなく容易」とは，(a) 変換元の言語資料に可能な限り人手で情報を付与しないこと，(b) 可能な限り簡単な操作で変換が実行できること，である。

以上の設計方針の下，『えだまめ』では，次の機能を実現した。

- 複数のプレーンテキスト，XML 文書から共有記述形式への変換
- データのファイル名，および，格納されているフォルダの構造を言語資料の属性として取得
- 変換したデータ用の『ひまわり』設定ファイルの自動生成

『えだまめ』は，Windows 上で動作し，変換のすべての操作を GUI (Graphical User Interface) を使って行うことができる。『えだまめ』の概観を図 4 に示す。本報告書では，変換の仕組みについて解説を行う。『えだまめ』の実際の利用方法については，『えだまめ』の公開ページを参照のこと[*6]。

4.1.2 変換の流れ

すでに述べたように，『えだまめ』は，複数のプレーンテキスト，もしくは，XML 文書から共有記述形式へ変換することができる。ここでは，プレーンテキストからの変換について説明することにする。

変換対象の言語資料は，「一定の規則」に基づいて，ファイルに名前を付け，「一定の規則」に基づいた階層構造のフォルダに格納する。この規則は，利用者が利用目的に基づいて決定する。例えば，図 5 は，検索結果

[*6] http://www.kokken.go.jp/lrc (「『ひまわり』支援ツール」) でマニュアルとともに，『えだまめ』本体を無償公開している。なお，2007-04-30 時点で公開しているのは，ver.1.2 であり，本報告書で述べている機能をすべて実現しているわけではない。

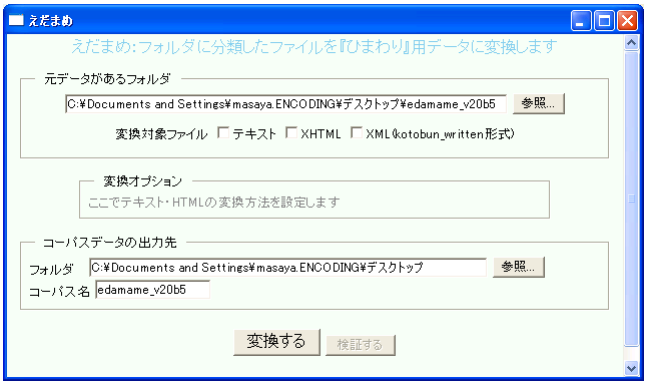

図 4　『えだまめ』の概観

として，時代区分と著者の情報を取得することを意図したものである。そのため，第 1 階層フォルダを時代区分 (例：近代，上代)，第 2 階層フォルダを著者名としている。なお，トップの階層の「テキストデータ」とは，『えだまめ』の変換対象データをまとめるためのフォルダである。

個々の言語資料は，この時代区分，著者名に基づいて，適切なフォルダに格納しておく。例えば，夏目漱石の「ぼっちゃん」であれば，ファイル名を「ぼっちゃん.txt」として，夏目漱石のフォルダに格納しておく。

以上の構成で言語資料を作成し，『えだまめ』で変換すると，次の処理が行われる。

- 文字コードの変換 (UTF-16 に変換される)
- ルビ，注タグの自動付与 (一定の形式でテキストが整形されている場合)
- 複数ファイルの結合 (一つのファイルにまとめられる)

作成された共有記述形式のデータを『ひまわり』にインストールすると，図 6 のような検索環境を実現できる。ディレクトリの構造やファイル名は，検索結果の「パス」欄に表示される。この情報を使えば，著者名や時代区分での絞込みやソートなどを『ひまわり』上で行うことができる。

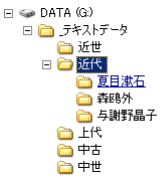

図 5　フォルダの階層構造

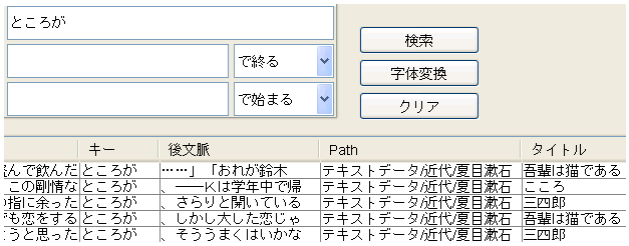

図 6　『えだまめ』で作成したデータの検索結果

4.2　『簡易変換ツール』

『簡易変換ツール』は，主として，XML で記述されない独自形式のテキストデータを共有記述形式に変換するツールである。『えだまめ』と異なり，利用には正規表現に関する知識が必要なので，一般の利用者を想定したツールではない。『簡易変換ツール』用の変換規則とともに，一般の利用者に配布して，無償で公開され

ていない言語資料に適用してもらう，という利用形態を想定している。本研究では，『ふみくら』「日本文学テキスト」「人情本」，および，「BTS による多言語話し言葉コーパス」をはじめとする四つの談話資料を共有記述形式に変換する際に利用している。

『簡易変換ツール』の動作を例に基づいて説明する。まず，次のような変換規則の集合を作成する。各行が一つの変換規則である。変換規則は，→の左右がそれぞれ変換前後の規則になっている。変換対象の言語資料に規則の左辺を適用し，マッチングが取れれば，右辺の文字列に変換する。左右の規則とも，正規表現[*7]で記述する。

```
^■ (.*)    →    <section>$1</section>
^■■ (.*)    →    <subsection>$1</subsection>
```

一つ目の規則は，行頭の■に後続する文字列を section タグでマークアップするものである。右辺の $1 は，左辺の () と対応しており，一つ目の () の中身で置き換えられる。同様に二つ目の規則は，行頭に■■がある場合，後続する文字列を subsection タグでマークアップするものである。

例えば，上記の規則を下図 (左) に適用すると，下図 (右) のように変換される。

■はじめに		<section>**はじめに**</section>
■■背景		<subsection>**背景**</subsection>
ここでは，背景について述べます。	→	**ここでは，背景について述べます。**
■■目的		<subsection>**目的**</subsection>

5　全文検索システム『ひまわり』の拡張

5.1　概要

本研究では，共有化された言語資料を全文検索する環境として，筆者らが開発した全文検索システム『ひまわり』を拡張する。本研究の目的である「言語資料の共有」ということを鑑み，拡張の内容として，次の２点を掲げる。

(1) 複数・大規模な言語資料への対応
(2) サーバ・クライアント型システムへの拡張

(1) では，複数の言語資料を共有して利用するということから，検索対象の資料を選択しやすくしたり，大規模なデータを高速に検索できるように拡張を行う。一方，(2) では，言語資料の大規模化に伴い，言語資料をサーバで管理できる仕組みを導入できるようにする。この機能を実現することにより，利用者は自分の計算機にたくさんの言語資料をインストールしなくても，クライアントシステムをインストールするだけで，サーバ上の言語資料を利用できるようになる。

[*7] 正規表現には，さまざまな「方言」が存在するが，『簡易変換ツール』では，Perl 言語の正規表現を採用している。

実際の拡張方法についての説明をする前に，全文検索システム『ひまわり』について簡単に説明しておく。『ひまわり』は，言語研究用に設計された全文検索システムであり，XML 形式の言語資料に対して全文検索を行い，KWIC 形式で表示したり，言語資料に付与された情報を検索することができる。図 7 は，『太陽コーパス』に対して，「研究」を全文検索した結果である。個々の検索結果に含まれる情報としては，検索文字列である「研究」に対する前後文脈のほか，「雑誌名」「年」「号」「題名」など，『太陽コーパス』に付与されているさまざまな研究用の情報を抽出することができる。さらに，KWIC よりも広い文脈を参照することを考え，図 7(右下) のように，記事全文を Web ブラウザに表示することができる。より詳細な情報は，別稿にゆずる。検索アルゴリズムの詳細については [13]，『ひまわり』を使った言語資料の作成方法については [14] を参照のこと。

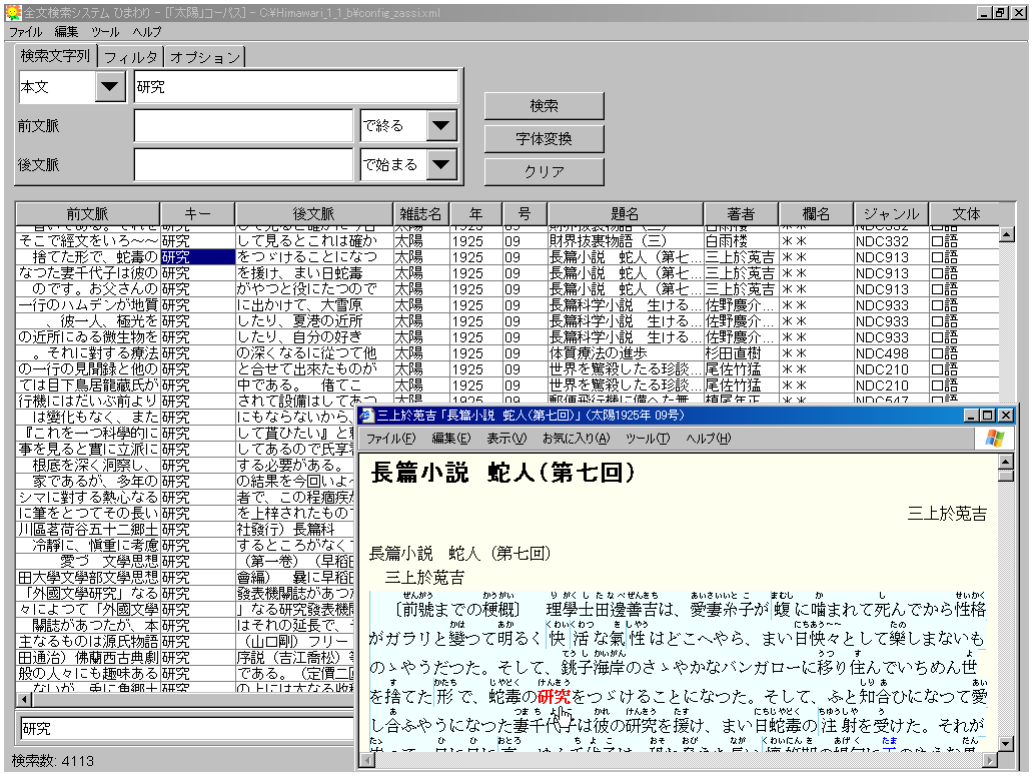

図 7　『太陽コーパス』への適用例 ([13] から引用)

5.2　設計と実現

5.2.1　複数・大規模な言語資料への対応

複数・大規模な言語資料への対応として，すでに一般に公開済みの『ひまわり』ver.1.2 に対して，次のような拡張を行った。

■検索アルゴリズムの見直し　新しい検索アルゴリズムでは，従来の検索アルゴリズムと比較して，使用するメモリ量を削減している。これは，より大きな言語資料を扱えるようになることを意味している。具体的なアルゴリズムの変更としては，索引の利用方法を図 8 のように変更している。

旧アルゴリズムでは，一つの検索結果を得るために，複数の索引を一度に利用する必要があった。例えば，行 (1) の検索結果を得るためには，「KWIC」「雑誌の情報」「記事個別の情報」の三つの索引を一度に利用する。同じ処理を行 (2),(3) に対しても実行する。それに対して，新アルゴリズムでは，まず，索引「KWIC」を利用して，列 (a) の情報だけを検索し，そのあと，索引「雑誌の情報」「記事個別の情報」を使って，順次列 (b)(c) の情報を検索する。こうすることにより，一度に利用する索引が一つになり，結果的にメモリの利用量が削減される。

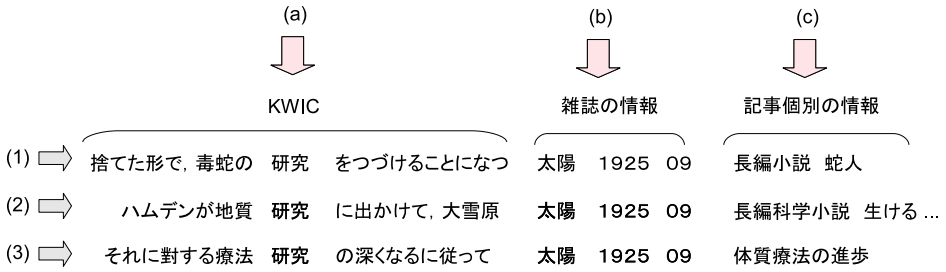

図 8　検索アルゴリズムの見直し

■用例抽出条件の追加　言語資料が大規模になると，検索結果の量が膨大になる場合がある。例えば，助詞の「が」を検索するために，『太陽コーパス』で「が」を全文検索すると，147352 件にもなる。したがって，スペックの低い計算機だとメモリ不足のため，検索できなかったり，使用に耐えうる時間で検索できないという問題が起こる。そこで，用例の抽出条件として，次の三つを追加することにした。図 9 は，用例抽出条件の指定用インターフェイスである。

(1) 抽出用例数を指定し，ランダムに抽出
(2) 用例数の計測のみを実施
(3) 検索数上限の設定

例えば，(2) は用例数のみを計測した上で，(1) を実施することにより，用例のランダムサンプリングを行うことができる。また，これまでの通常の検索に対しても (3) で検索数上限を指定することができる。

■コーパス選択用インターフェイスの実現　従来の『ひまわり』ver.1.2 でも複数のコーパスを選択的に利用することは可能であったが，『ひまわり』の設定ファイルを直接編集しなければならないため，一般の利用者には利用が困難だった。そこで，図 10 のようなインターフェイスを設けて，容易に検索対象の言語資料を選択できるようにした。図 10 では，新聞を 3 ヵ月を単位に一つの言語資料としているが，「追加」「除外」ボタンを押すことにより，検索対象の言語資料を取捨選択することができる。

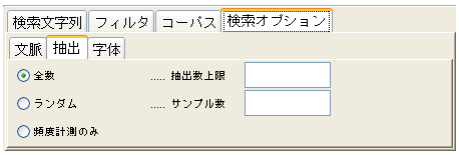

図 9　抽出条件の指定

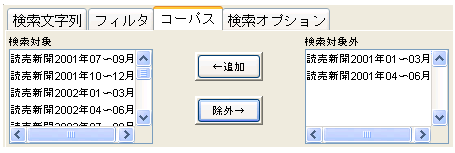

図 10　言語資料の選択

■複数の検索結果テーブルの実現　言語資料を利用して，語の用例分析を行う場合，複数の語の用例を一度に比較したい場合がある。例えば，作者ごとの用法の違いなどの分析などである。このような要求は，利用できる言語資料が増えると，増大してくるものと思われる。

そこで，複数の検索結果テーブル (図 7 では，検索文字列「研究」の検索結果の一覧が表示されている部分) を利用できるように『ひまわり』を拡張した。図 11 が実現した結果である。例えば，図 11(左) のように，タブを追加し，新たな検索結果テーブルを作成する。また，図 11(右) のように，テーブルを削除することもできる。このように，個々の検索結果テーブルにタブをつけて，複数の検索結果を格納し，それぞれを比較しやすくした。

図 11　複数の検索結果テーブル

5.2.2　サーバ・クライアント型システムの実現

サーバ・クライアント型システムを実現するにあたって，次の設計方針を立てた。

- さまざまな種類のクライアントからの検索要求に答えられるようにする。例えば，『ひまわり』や一般の Web ブラウザを介しての検索要求である。
- 本研究で実現するクライアントは，可能な限り，現在の『ひまわり』と同等の機能が実現できるようにする。
- 一般の利用者が手軽にサーバを設置できるようにする。例えば，自宅や大学の研究室で，複数の利用者が検索を行うような状況を想定する。

そこで，図 12 の構成でサーバ，クライアントシステムを構築することにした。

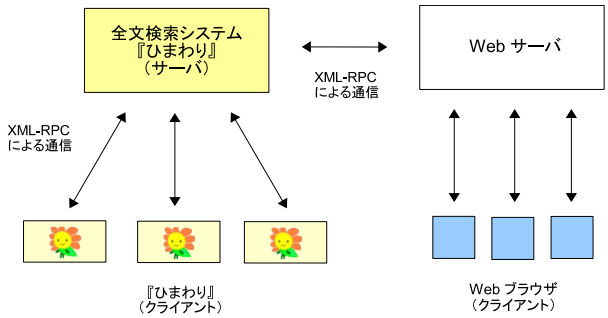

図 12　サーバ・クライアント型システムの構成

■『ひまわり』サーバ　まず，『ひまわり』サーバは，クライアントシステムからの検索要求を受け，検索結果などを返す役目を果たす。『ひまわり』サーバは，XML-RPC (XML Remote Procedure Call)[15] と呼ばれる仕組みを使って，クライアントと検索要求，検索結果などのやりとりを行う。XML-RPC を使う利点は，次

の二つである。一つは，XML-RPC は仕様が単純であるとともに，Perl，Java，C など多くのプログラミング言語で利用することが可能なので，クライアントシステムを構築しやすいことである。もう一つは，HTTP 上でサーバ，クライアント間の通信が行われるので，多くのネットワークで利用できることである[*8]。

『ひまわり』サーバの実現にあたっては，従来の『ひまわり』にサーバ機能を付加する形で行った。『ひまわり』サーバは従来の『ひまわり』の言語資料をそのまま利用することができる。また，『ひまわり』サーバを起動するには，従来の『ひまわり』を「サーバモード」で起動するだけでよい[*9]。以上のように，『ひまわり』サーバを設置することは容易であり，一般の利用者でも自宅や研究室で『ひまわり』サーバを利用できると予想される。

現在のところクライアントから受け付けている要求は，次の三つである。

- サーバの初期化
- 検索結果の出力形式に関する問い合わせ
- 検索要求

クライアント側からの一般的な通信手順は，次のとおりである。まず，「サーバの初期化」を実行する。次に，「検索結果の出力形式に関する問い合わせ」を行い，検索結果にどのような情報が含まれているか問い合わせる（例えば，「検索結果の第 1 フィールド目は「前文脈」の情報が格納されている」など）。最後に，利用者が指定した検索条件でサーバに検索要求を行い，サーバの返した検索結果を表示する。

■クライアント　クライアントシステムは，図 12 のように，2 通りの形態を想定する。一つは，従来の『ひまわり』から『ひまわり』サーバにアクセスする方法である。もう一つは，Web サーバを介して，検索を実行するタイプのクライアントである。後者は，図のように，利用者が Web ブラウザを使って，Web サーバにアクセスする。さらに，Web サーバがサーバ版『ひまわり』に対するクライアントシステムを起動し，検索を実行する。そして，検索結果をユーザの Web ブラウザに返す。本研究では時間の都合上，『ひまわり』クライアントを実際に実現することにした。

『ひまわり』クライアントは，従来の『ひまわり』にサーバとの通信機能を付加した。サーバに接続するには，拡張された『ひまわり』の設定ファイルに，次の 3 行目のような記述[*10]を追加するだけでよい。

```
<li name="「太陽」" path="Corpora/Zassi/Taiyo/corpus" />
<li name="女性雑誌" path="Corpora/Zassi/Josei/corpus" />
<li name="『テストコーパス』" path="http://www2.kokken.go.jp:8085" />
```

『ひまわり』クライアントは，指定されたサーバ (上の例の場合，www2.kokken.go.jp) に接続し，XML-RPC を用いて，検索要求，検索結果の取得などを実行する。なお，上から二つの設定は，『ひまわり』クライアント本体に格納されている言語資料に関する指定である。このように，すでに利用している言語資料と混在させて検索を行うことも可能である。

[*8] セキュリティ上の制約から HTTP 以外のプロトコルの利用に制限をかけているネットワークも存在する。

[*9] 「サーバモード」で起動するには，『ひまわり』の起動時オプションで指定する。ただし，WindowsXP 上で『ひまわり』サーバを実行する場合，WindowsXP の「ファイアウォール機能」を解除する必要がある。

[*10] この記述は，あくまで例である。このサーバに接続しても『ひまわり』サーバは起動していない。

6　全体評価

ここでは，本研究全体の評価として，三つの観点から評価を行うとともに，今後の課題について述べる。

■共有記述形式の記述能力について　今回設計した共有記述形式のタグ数は，書き言葉用 24 種類，話し言葉用 15 種類と，十分にシンプルな仕様になったと考えている。また，3.4 節で示したように，書き言葉の言語資料合計 2635 作品，話し言葉の言語資料 4 点を共有記述形式に変換し，『ひまわり』での動作を確認した。したがって，記述能力についても目標を達成したと考える。

今後の課題としては，実際の研究者が共有記述形式で言語資料を作成することができるか，確認することである。

■共有記述形式への変換のしやすさについて　本研究では，変換を支援するツールとして，『えだまめ』『簡易変換ツール』の 2 種類を作成した。このうち，変換の簡便性に焦点をあてて設計した『えだまめ』は，プレーンテキストからマウス操作だけで共有記述形式のデータを作成することができる。したがって，変換のしやすさという点では目標を達成できたと考える。

今後の課題としては，『えだまめ』の変換能力を向上させることである。現状では，プレーンテキストから共有記述形式のすべてのタグを生成できるわけではなく，ルビや注記のタグに限られている。そこで，今後プレーンテキストからより多くのタグを容易に生成できる仕組みを考える必要がある。

■言語資料の利用しやすさについて　既存の全文検索システム『ひまわり』を拡張し，(1) 複数・大規模なコーパスへ対応，(2) サーバ・クライアント型のシステムへの拡張を実現した。これにより，一般の利用者が複数・大規模なコーパスを利用しやすい環境を実現できたと考える。

今後の課題としては，サーバ版『ひまわり』の機能を高めることである。現在のところ，基本的な検索を実行することはできるが，改良すべき点は多い。例えば，サーバからの部分的な検索結果の取得を実現することである。現状では，大量の検索結果が得られた場合，すべての結果をクライアントに転送するため，検索速度が遅くなる。そこで，先頭 100 件というように，部分的な検索結果を取得できるようにする。

7　おわりに

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を想定し，電子化された言語資料の共有，利用を支援する環境を構築した。具体的には，(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，である。本研究により，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構築することができた。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。

謝辞

共有記述形式用の資料を提供してくださるとともに，Web 上への公開を許可してくださった千葉大学文学部の岡部嘉幸氏，高木 元氏，大阪大学の岡島昭浩氏に深く感謝する。言語資料の調査，共有記述形式の仕様

の作成，データの変換など，本研究のあらゆる場面で多大な協力をしてくださった上野左絵氏，藤本三輪氏に感謝する。

参考文献

[1] 安永尚志： 国文学研究とコンピュータ，勉誠社 (1998)
[2] 原正一郎，安永尚志： 文学研究のためのデータベースシステムの諸問題 —文学データ共有のための標準化 —，日本語学 Vol.20 No.13，pp.48–60 (2001)
[3] Text Encoding Initiative, The XML Version of the TEI Guidelines, http://www.tei-c.org/P4X/index.html
[4] Corpus Encoding Standard, http://www.cs.vassar.edu/CES/
[5] 宇佐美まゆみ（東京外国語大学）監修：『BTS による多言語話し言葉コーパス－日本語会話 1（日本語母語話者同士の会話)』
[6] 宇佐美まゆみ（東京外国語大学）監修：『BTS による多言語話し言葉コーパス－日本語会話 2（日本人と学習者の会話)』
[7] 宇佐美まゆみ： 改訂版：基本的な文字化の原則（Basic Transcription System for Japanese: BTSJ)」『多文化共生社会における異文化コミュニケーション教育のための基礎的研究』(科学研究費補助金基盤研究 (C) 2 ：研究代表者 宇佐美まゆみ）研究成果報告書
[8] 現代日本語研究会編： 女性のことば・職場編，ひつじ書房 (1997)
[9] 現代日本語研究会編： 男性のことば・職場編，ひつじ書房 (2002)
[10] 遠藤織枝 他： 戦時中の話しことば　ラジオドラマの台本から，ひつじ書房 (2004)
[11] 野口英二：『インターネット図書館青空文庫』，はる書房 (2005)
[12] 田中牧郎，山口昌也，吉田谷幸宏，小木曽智信，近藤明日子 (国立国語研究所 編)：太陽コーパス 雑誌『太陽』日本語データベース，博文館新社 (2005)
[13] 山口昌也, 田中牧郎：構造化された言語資料に対する全文検索システムの設計と実現，自然言語処理 vol.12, No.4, pp.55–77 (2005)
[14] 山口昌也：全文検索システム『ひまわり』を利用した言語資料検索環境の構築手法，日本語科学 vol.21, pp.111-123 (2007)
[15] http://www.xmlrpc.com/spec