全文検索システム『ひまわり』を用いた既存言語資料の活用方法の検討

山口昌也 (国立国語研究所言語資源研究系) †

Exploitation of Existing Language Resources by Full-Text Search System “Himawari”

Masaya YAMAGUCHI (Dept. Corpus Studies, NINJAL)

要旨

本稿では，筆者が開発している全文検索システム『ひまわり』を対象として，既存の言語資料を活用する方法を検討した。既存の言語資料を利用するための検索ツールを考える場合，言語資料の記述形式に対応できるよう機能を実装するとともに，言語資料の再配布可能性や規模などの性質に応じた運用方法を検討する必要がある。本稿では，三つの言語資料を全文検索システム『ひまわり』にインポートし，他のユーザと共有可能な状態にする過程をとおして，(1) 多様なデータ形式に対するインポート能力を確認した，(2) 大規模なデータに対応するためにサブコーパス単位のインポート機能を実現した，(3) 利用条件に応じた配布形態に対応するために，配布資料のパッケージ化を図った。

1　はじめに

本稿では，筆者が開発している全文検索システム『ひまわり』(山口，田中 2005)[1] を用いて，既存の言語資料を活用する方法を検討する。現在，言語資料の活用を支援するためのシステムが数多く提案されており，Web ベースのコーパス検索システム (今井ら 2013，小木曽ら 2011 など)，コーパス管理機能を備えた検索システム (松本ら 2006 など)，高度な分析機能を備えた検索システム (樋口 2003 など) などが利用可能になっている。これらの検索システムに対して,『ひまわり』は，(1)XML により記述された多様な形式の言語資料の全文検索・閲覧，(2) 多様な形式の言語資料のインポート機能などの特徴を持っている。本稿では，特に，(2) の特徴を活かして，既存の言語資料を『ひまわり』で活用する方法を考える。

山口 (2013) では，既存の言語資料を『ひまわり』用のデータ形式に変換するための方法を検討・実装した。しかし，個人が作成するような小規模な言語資料を想定しており，言語資料の規模や，他の研究者と共有する際の問題については，十分考慮していなかった。また，現在のところ，実際の言語資料への適用は 1 例 (『青空文庫』XHTML 版) のみであり，特に独自形式のテキストデータをどの程度インポートできるのか，検証が十分でなかった。

そこで，本稿では，(1) 既存の言語資料を『ひまわり』にインポートし，他の研究者と共有可能な状態にするよう試みる，(2) その過程で発生する問題を明らかにする，(3) 解決するための仕組みを『ひまわり』に実装する，という手順で既存の言語資料を活用する方法を検討した。

2　既存の言語資料の活用

2.1　対象とする言語資料

今回使用した言語資料は,『日本語話し言葉コーパス』[2](以後，CSJ),『CD- 毎日新聞データ集』[3](以後,「毎日新聞」)，米国議会図書館蔵『源氏物語』[4](以後,「源氏物語」) の 3 種類である。これらを『ひまわり』にインポートし，他の研究者に配布する場合，どのような問題が発生するかを，(a) データ形式，(b) データの規模，(c) 利用条件，の三つの観点から考えてみる。上記の言語資料は，これら三つの観点の上で特徴がでるように選択した。各言語資料の内訳を表 1 に示す。

†http://www2.ninjal.ac.jp/masaya
[1]http://www2.ninjal.ac.jp/lrc/
[2]http://www.ninjal.ac.jp/corpus_center/csj/
[3]http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html
[4]http://textdb01.ninjal.ac.jp/LCgenji/

表 1: 対象とする言語資料の内訳

言語資料	データ形式	規模	利用条件
CSJ	XML	2.9GB	有償・再配布不可
毎日新聞	独自形式テキスト	300〜600MB(1 年分)	有償・再配布不可
源氏物語	独自形式テキスト	2.8MB	無償・再配布可能

2.2　活用する際の問題と対策

2.2.1　データ形式

言語資料は，個々の形式に従って記述されている。そのため，それらを解釈して『ひまわり』で検索するための形式 (以後,『ひまわり』形式) に変換する必要がある。

この問題に対して，山口 (2013) では，言語資料ごとに変換規則を用意することにより，多様な形式の言語資料のインポートに対応している。想定するデータ形式は，独自形式のテキスト，および，XML(HTML は内部的に XHTML へ変換) である。変換規則は，独自形式のテキストの場合は正規表現による文字列置換規則 (詳細は，後述)，XML の場合は XSLT スタイルシートにより記述する。前述のとおり，この機能はすでに公開中の『ひまわり』に実装されているため，本稿では実際の言語資料に対する検証のみを行う。

2.2.2　データの規模

一般的に，検索システムで言語資料を利用する場合，扱えるデータの規模に制限がある。これは，検索システムのハードウェア上の制限のほか，規模を大きくしすぎると実用的な検索時間で結果を得られないなど，利用上の制約が存在するからである。『ひまわり』の場合，広範な PC での動作を考慮すると，単一の『ひまわり』形式データの上限は，おおむね 150MB 程度である。

データの規模の問題に対する対策としては，サブコーパスに分割し，それぞれ独立した『ひまわり』形式データして管理することが考えられる。この方法は，CSJ のように，複数のサブコーパスからなるコーパスにとっては自然な対応であろう。また,「毎日新聞」のように，1 年分が数百 MB になるテキストデータベースの場合も，単年ごとにサブコーパスとして管理するほうが使い勝手がよいと考えられる。例えば，検索時間や検索量を試しつつ，必要に応じて，検索対象を調節するといったことが可能になる。

『ひまわり』では，複数の『ひまわり』形式データを個別に検索し，結果を統合する機能がすでに実装されている。ただし，言語資料のインポート時は，常に単一の『ひまわり』形式データとなる。そこで，インポートするファイル群のディレクトリ構造に基づいて，複数の『ひまわり』形式データを生成するように，インポート機能を拡張する。具体的には，インポート対象のファイルを収録したディレクトリをルートディレクトし，その直下のディレクトリごとに『ひまわり』形式データを構築するようにする。

2.2.3　データの利用条件

今回対象とする言語資料の利用条件は，２種類に分けられる。一つは,「源氏物語」のような「改変・再配布可能」である。この利用条件であれば，誰でも変換後の言語資料を配布することができる。一方，CSJ と「毎日新聞」のように「有償・再配布不可」の場合，言語資料の権利者でなければ，変換結果の言語資料を配布することができない。

後者のような言語資料を多くの利用者が『ひまわり』から利用できるようにする対策としては，インポート用の変換規則を配布し，利用者自身がインポートすることである。ただし，現状の『ひまわり』ではインポートはできても,『ひまわり』の設定ファイル自体は，汎用の設定ファイルが自動生成されるだけである。そのため，個々の言語資料が持つ特徴を活かした検索や閲覧がしづらい。そこ

で，インポート時に，指定された『ひまわり』用の設定ファイルが同時にインストールされるように，変換規則と設定ファイルのパッケージ化を図った。

また，これに合わせて，変換後の言語資料を配布する場合のインストールもパッケージ化した。具体的には，従来，インストールするファイルのコピーをユーザが手動で行わなければならなかったが，自動的にインストールできるようにした。

3　インポートの詳細

3.1　CSJ

CSJ は XML で記述されたコーパスであり，利用条件は有償・再配布不可である。1 講演 1 ファイルで合計 3302 ファイルで構成される。有償・再配布不可であることから，言語資料自体の配布を行うのではなく，変換規則を配布する形態を考える。

CSJ はさまざまな形式で配布されているが，本稿では CSJ の「XML 文書」形式のデータ [5] を用いた。この形式のデータには，CSJ に収録されているほとんどの付与情報を含んでいる。ただ，今回は，形態素解析済みのテキストとして利用することを目的にし，「XML 文書」から転記テキスト，および，「短単位・長単位」の情報を抽出して，『ひまわり』形式に変換するようにする。XML で記述されているので，変換は XSLT により行う。変換用のスタイルシート自体は，すでに一般公開済みのものを用いた。詳細は，公開ページ [6] を参照されたい。

CSJ の規模は全体で約 2.9GB あることから，2.2.2 節で述べたように，複数のサブコーパスへ分割する必要がある。分割の単位は，音声タイプ (例：学会講演，模擬講演など) と音声タイプの詳細情報 (例：学会の別，模擬講演テーマの別など)，人手解析・自動解析の別を基準として，合計 16 個に分割した。なお，分割時は 16 個のフォルダに 3302 個のファイルを移動することになり，ユーザの手間が大きい。そのため，ファイル振分け用のシェルスクリプトを用意することにより対処する。

3.2　「源氏物語」

「源氏物語」は，独自形式で記述されたテキストデータであり，1 冊 1 ファイルで計 54 ファイルから構成される。利用条件は無償・再配布可であることから，変換後の『ひまわり』形式のデータを配布する形態を考える。

「源氏物語」のテキストデータの例を図 1 に示す。全体的な構造としては，資料のタイトル，資料説明，本文，作成者情報，本文修正情報から構成されている。本文中の付与情報は，ページ情報，和歌の範囲の２種類である。他の２資料と比較して，特徴的なのは，ページや行の区切りの情報が改行文字で記述されていることである。このようなテキストデータを全文検索する場合，行やページをまたぐ語や表現が検索できなくなるため，『ひまわり』形式データへの変換時に対策が必要である。

『ひまわり』形式データへの変換に際しては，(a) 本文部分の全文検索が確実にできること，(b) 検索文字列が含まれる作品タイトル，ページ位置番号を取得できることを目標として，次のような文字列置換規則を作成した。規則数は 9 個である。結果の一部を図 3 に示す。なお，図中の「→」の左辺は，正規表現で記述された変換対象の文字列，右辺が変換結果の文字列である。

- 1 冊を genji 要素とした。冊のタイトルを genji 要素の title1, title2 属性 (それぞれ漢字表記，原文表記) に変換した。[規則 1]
- ページ区切り位置は page 要素 (空要素) とし，ページ番号はその no 属性に記述した。[規則 2]
- 行区切りのための改行は，改行を表す br 要素 (空要素) とした。また，page 要素前後の改行文字はすべて削除し，1 冊が 1 文字列になるように連結した。これにより，行，ページにまたがる語，表現の検索ができるようになる。[規則 3，5]

[5] http://www.ninjal.ac.jp/corpus_center/csj/manu-f/xml.pdf
[6] http://www2.ninjal.ac.jp/lrc/ 中の『ひまわり』のホームページから，「『日本語話し言葉コーパス』を『ひまわり』で利用する方法 」を参照。

```
米国議会図書館蔵『源氏物語』
桐壺
--------------------------------------------------------------------------------
記号の説明
1. くの字点は／＼で表す。
2. 和歌は「」で括る。
      :
--------------------------------------------------------------------------------

きりつほ

（１オ）
いつれの御時にか女御更衣あまたさふらひ給けるなかに
いとやむことなききはにはあらぬかすくれてときめき
給ふありけりもとより我はと思ひあかりたまへる御かた／＼
めさましき物におとしめそねみ給ふおなし程それより
けらうの更衣たちはましてやすからすあさ夕のみや
つかへにつけても人の心をのみうこかしうらみをおふつもり
にやありけんいとあつしくなりゆき物心ほそけにさとかちに
なるをいよ／＼あかすあはれなる物におほして人のそしりをも
```

図 1: 「源氏物語」のテキストデータ例

```
<genji main_title="米国議会図書館蔵『源氏物語』" title1="桐壺" title2="きりつほ" >
<comment>
--------------------------------------------------------------------------------
記号の説明
1. くの字点は／＼で表す。
2. 和歌は「」で括る。
     :
--------------------------------------------------------------------------------
</comment>
<body>
<page no="０１オ" />いつれの御時にか女御更衣あまたさふらひ給けるなかに<br />いとやむことなききはにはあらぬかすくれてときめき<br />給ふありけりもとより我はと思ひあかりたまへる御かた／＼<br />めさましき物におとしめそねみ給ふおなし程それより<br />けらうの更衣たちはましてやすからすあさ夕のみや<br />つかへにつけても人の心をのみうこかしうらみをおふつもり<br />にやありけんいとあつしくなりゆき物心ほそけにさとかちに<br />なるをいよ／＼あかすあはれなる物におほして人のそしりをも<br /><page no="０１ウ" />えは〻からせたまはす世のためしにもなりぬへき御もてなし也かん<br />たち
```

図 2: 「源氏物語」の変換例

```
# 規則１　genji 要素（開始）と冒頭の comment 要素
^(?s)(.+?)\n(.+?)\n(.*?---+\n.+?---+\n)\n*(.+?)\n
 → <genji main_title="$1" title1="$2" title2="$4" >\n<comment>\n$3</comment>\n<body>\n
# 規則２　page 要素
(?s)（(.{2,3}?)）\n(.+?)\n\n　→ <page no="$1" />$2\n</dummy>
# 規則３　page 要素内の改行を br 要素で置換する
(?s)(?<=<page [^<]{0,30}?>[^<]{0,1000}?)\n(?!</dummy>)　→ <br />
# 規則４　和歌部分を waka 要素とする
(　+「.+?」)　→　<waka>$1</waka>
# 規則５　ページ末尾を br にする
\n+</dummy>　→ <br />
# 規則６　資料末尾のコメントを comment 要素とし，gengi 要素の閉じタグを出力する
(?s)\n(---+\n[^<]+)$　→ \n</body>\n<comment>\n$1\n</comment>\n</genji>\n
```

図 3: 「源氏物語」用の文字列置換規則 (一部)

- 和歌を waka 要素とした。[規則 4]
- 本文を body 要素とし，その前後にある資料の説明，本文，作成者情報，本文修正情報などは comment 要素とした。全文検索時は通常 body 要素内のみを検索するように設定することにより，本文のみを検索対象とすることができる。[規則 1 ， 6]

なお，それぞれの規則は，入力テキストファイルに対して，上から順に一つづつ適用される。「源氏物語」の場合， 1 ファイル 1 冊なので，54 回分独立に規則が適用され，結果は一つの『ひまわり』形式データに合併される。注意すべきことは，規則の適用が 1 行ごとではなく， 1 ファイルを 1 文字列とした状態で行うことである。これにより，複数の行にまたがった文字列に対する置換ができるようになっている。

3.3　毎日新聞

「毎日新聞」は，独自形式で記述されたテキストデータであり，1 年分の記事が 1 ， 2 ファイルにまとめられている。「毎日新聞」の場合，1991 年から 2013 年までの 23 年分のデータが販売されている。利用条件は有償，再配布不可であることから，文字列置換規則を配布する形態を考える。

「毎日新聞」のテキストデータの例を図 4 に示す (日外アソシエーツの Web ページ [7] から引用)。データは，各行がデータ種別とその内容になっている構造である。例えば，「＼Ｔ１＼」で始まる行は記事見出し，「＼Ｔ２＼」で始まる行は記事全文である。記事中に付与情報は記述されていないが，「＼Ｔ１＼」のような形式で記事の情報が記述されている。

```
＼ＩＤ＼００００００１０
＼Ｃ０＼０１０１０１００１
＼ＡＤ＼０１
＼ＡＥ＼Ｎ
＼ＡＦ＼０１０１０１Ｍ０１
＼Ｔ１＼［余録］カンボジアの太陽は日本で見るより大きく見える…
＼Ｓ１＼　　’０１．１．１　朝刊　１頁　写図無　（全７３５文字）
＼Ｓ２＼　カンボジアの太陽は日本で見るより大きく見える。そのせいか１２月でも暑い。日が落ちて午後６時前に空は濃紺に染まり、やがて黒一色になる▲アンコールワット近郊のモンドルバイ村。対人地雷の犠牲になった人々の住む障害者村だ。電気はない。月がないときは真の闇だ。光を奪われた村民のために、ＮＧＯ「国際人権ネットワーク」代表、緒方由美子さんが使いかけのろうそくを集めて贈ることにした話は昨年５月、この欄でご紹介した▲「新聞で見て、雨戸をしめ、真っ暗な中でろうそくを１本つけました。何とホッとすることか。心が安らぎます。私たちも協力させて下さい」と電話をかけてきたお年寄りもいる。こうして日比谷花壇から結婚式のろうそく４０００本、記事を読んだ人、伝え聞いた人から４０００本が集まった▲昨年１２月１日、大勢の人の思いを乗せて、モンドルバイ村６００世帯にろうそくが配られた。
```

図 4: 「毎日新聞」のテキストデータ例

『ひまわり』形式データへの変換では，記事に付与されている情報を検索時に取得できるようにすることを目標にした。図 5 に変換結果を示す。付与情報は，記事を表す at 要素の属性として，発行年・月・日，面種，タイトル，朝夕刊の別，記事文字数などを取り込んだ。文字列置換規則数は，60 個である。「源氏物語」に比べて，規則が多くなったのは，コード化されている面種をデコードしたり，at 要素の数値属性値を半角文字列に統一するなどの置換を行っているためである。

前述のとおり，「毎日新聞」は 1 ファイル (半年，もしくは， 1 年分) のサイズが 300〜600MB と巨大なので， 1 ファイル 1 サブコーパスとして，複数のサブコーパスに分割する。今回は， 3 年分 (4 ファイル) をインポート対象とした。

3.4　インポート結果

表 2 にインポート結果を示す。インポートに使用した PC のスペックは，CPU: Intel Corei5 2.53GHz，メモリ:4GB，OS: MacOS 10.9.3 である。変換自体はどの言語資料も成功した。

[7] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html

<at y="2001" m="01" d="01" g="1 面" t="［余録］カンボジアの太陽は日本で見るより大きく見える…" p="朝刊" c="735">
　カンボジアの太陽は日本で見るより大きく見える。そのせいか１２月でも暑い。日が落ちて午後６時前に空は濃紺に染まり、やがて黒一色になる▲アンコールワット近郊のモンドルバイ村。対人地雷の犠牲になった人々の住む障害者村だ。電気はない。月がないときは真の闇だ。光を奪われた村民のために、ＮＧＯ「国際人権ネットワーク」代表、緒方由美子さんが使いかけのろうそくを集めて贈ることにした話は昨年５月、この欄でご紹介した▲「新聞で見て、雨戸をしめ、真っ暗な中でろうそくを１本つけました。何とホッとすることか。心が安らぎます。私たちも協力させて下さい」と電話をかけてきたお年寄りもいる。こうして日比谷花壇から結婚式のろうそく４０００本、記事を読んだ人、伝え聞いた人から４０００本が集まった▲昨年１２月１日、大勢の人の思いを乗せて、モンドルバイ村６００世帯にろうそくが配られた。

図 5: 「毎日新聞」の変換例

最後に，2.1 節で示した三つの観点から結果を評価する。まず，データ形式については，懸案だった独自形式のテキストに対しても，目標に沿った変換を行うことができた。プログラミング言語を用いず，正規表現の置換規則のみで変換できるので，利用者の学習コストを抑えることができると思われる。

データの規模に対しても，サブコーパスのインポート機能を実装したことにより，大規模なデータのインポートが容易になった。ただし，サブコーパスの単位を決める際のファイルの振り分けにシェルスクリプトが必要になるなど課題もある。

利用条件の問題については，利用条件に応じた資料配布形態 (変換規則の配布，変換後の言語資料の配布) に対応するため，配布資料のパッケージ化を図った。今回の三つの資料についての動作は確認している。

表 2: インポート結果

言語資料	入力データサイズ	結果データサイズ	変換時間
CSJ	2.9GB	1.9GB	72min
毎日新聞	1.5GB	0.98GB	25min
源氏物語	2.8MB	2.2MB	0.23min

4　終わりに

本稿では，三つの言語資料を全文検索システム『ひまわり』にインポートし，他のユーザと共有可能な状態にする過程をとおして，(1) 多様なデータ形式に対するインポート能力を確認した，(2) 大規模なデータに対応するためにサブコーパス単位のインポート機能を実現した，(3) 利用条件に応じた配布形態に対応するために，配布資料のパッケージ化を図った。

参考文献

山口昌也, 田中牧郎 (2005) 「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

今井新悟，赤瀬川史朗，プラシャント・パルデシ (2013) 筑波ウェブコーパス検索ツール NLT の開発，第３回コーパス日本語学ワークショップ予稿集，pp.199–206

小木曽智信，中村壮範，鈴木泰山，八木豊，山崎誠，前川喜久雄 (2011) 「コーパス検索システム「中納言」デモンストレーション」，日本語コーパス完成記念講演会予稿集，pp.43–46

樋口耕一 (2003)「コンピュータ・コーディングの実践 —漱石『こころ』を用いたチュートリアル—」，年報人間科学 24，pp.193–214

松本裕治，浅原正幸，橋本喜代太，投野由紀夫，大谷朗，森田敏生 (2006) 「タグ付きコーパス管理/検索ツール『茶器』」，言語処理学会第 12 回年次大会論文集，pp.460–463