個人用コーパスの作成とアノテーションを支援する環境の実現

山口昌也 (国立国語研究所言語資源研究系) †

Implimentation of an Environment for Personal Corpus Construction and Annotation

Masaya YAMAGUCHI (Dept. Corpus Studies, NINJAL)

1 はじめに

本稿では，個人の研究者がコーパスを構築し，アノテーションするのを支援するための環境について述べる。本環境は，全文検索システム『ひまわり』(山口・田中 2005) を機能拡張することにより実現する。

現在，Web データをはじめとして，大規模な電子的テキストを容易に入手できるため，個人の研究者でもコーパスを構築できるようになっている。また，コーパスを構築したり，活用したりするためのシステムとしても，Web ベースのコーパス検索ツール『中納言』(小木曽ら 2011)，コーパス管理ツール『茶器』(松本ら 2006)，テキストマイニングツール KHCorder(樋口 2003) など，様々なツールが利用できる。ただし，収集したデータを用いて，研究者自らがコーパスを構築し，言語研究用に利用することを考えた場合，いくつかの技術上の障害がある。本稿では，コーパス構築時のデータの統一性と，コーパス構築後のアノテーションに焦点を絞って，研究者を支援する方法を考える。

2 支援の方法

2.1 全体的な支援の流れ

図 1 に全体的な支援の流れを示す。提案する支援方法は，図の左側のコーパス構築時の支援（図 1 左）と，構築後の追加的なアノテーション支援（図 1 右）に分けられる。

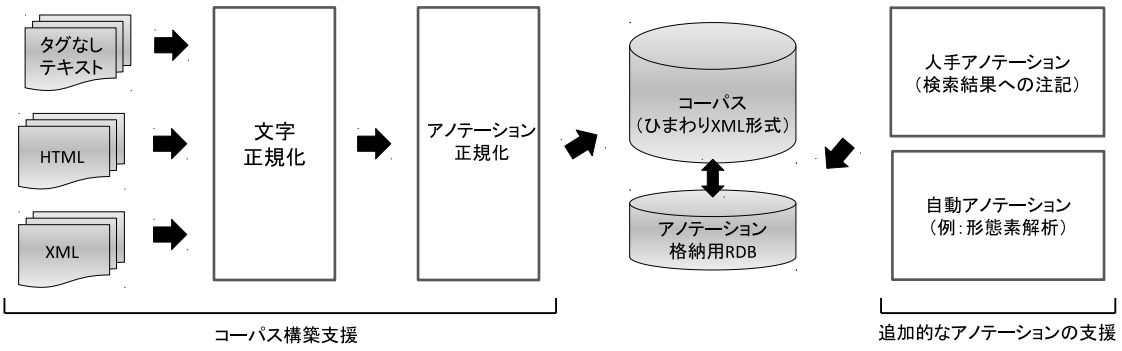

図 1: 全体的な支援の流れ

コーパスの構築時の支援としては，複数の電子テキストを統合する際の (1) 文字の正規化，(2) アノテーションの正規化，の 2 点を行う。原資料とするテキストの形式は，タグなしテキスト，HTML，XML とする。コーパス構築支援の結果，生成されるコーパスは，統一された XML 形式（ひまわり XML 形式）のファイルとなる。

また，コーパス構築後の支援として，追加的なアノテーションの支援を行う。支援の方法は，2 種類ある。一つは，形態素解析システムなどの自然言語処理システムの解析結果のアノテーション支援である。もう一つは，検索結果に注記などを加える，人手のアノテーションの支援である。

†http://www2.ninjal.ac.jp/masaya

2.2 コーパス構築支援

2.2.1 文字の正規化

いわゆる半角・全角数字のように，同一の文字が電子的に別の文字として扱われると，検索漏れの原因となる。ここでは，文字の正規化として，符号化方式の正規化と文字コードポイントの正規化を行う。

まず，符号化方式の正規化についてである。現在，電子化テキストの符号化には，Shift JIS，UTF-8，EUC など，多くの符号化方式が用いられる。これらをひまわり XML 形式データで用いている UTF-16 に統一する。なお，この際，原資料の符号化方式は，自動推定している。

符号化方式の正規化に加えて，文字のコードポイントの正規化を行う。具体的には，複数のコードポイントに割り当てられている文字や，合成を含む文字を単一のコードポイントに統合する。前者の例として，いわゆる半角，全角の英数字を挙げる。後者の例としては，テ＋゛⇒デのような，半角の仮名文字と濁点の合成がある。

コードポイントの正規化の方法としては，二つのオプションを用意している。一つは，変換テーブルを用いるもので，拡張版の『ひまわり』には，半角英数文字（JIS X 0201 のラテン文字用図形文字）から全角文字への変換テーブルが付属している。もう一つは，Unicode の規格として規定されている正規化形式 (NFKC, Normalization Form Compatibility Composition) に基づく正規化 (Davis and Whistler 2012) である。

2.2.2 アノテーション形式の正規化

前述のとおり，原資料のファイル形式は，タグなしテキスト，HTML，XML の 3 種類である。『ひまわり』は，個々の形式ごとに変換規則を定め，それぞれの形式からひまわり XML 形式に変換する。ファイル形式ごとに変換方法は，次のとおりである。なお，アノテーション形式の正規化については，従来から公開している，ひまわり XML 形式データの作成支援ツール『えだまめ』の機能に拡張を加えたものとなっている。

タグなしテキスト　タグなしテキストは，一般的な統一規格に基づいたタグでアノテーションされていないが，独自形式のタグでアノテーションされている場合がある。例えば，青空文庫の「テキストファイル」形式では，ルビが「五月雨《さみだれ》」のような形式で記述されている。

このような独自形式のタグを，利用者が定義した文字列置換規則により，XML タグへ変換する。文字列置換規則は，指定された正規表現にマッチした文字列を指定された文字列に置き換える。

HTML　HTML でアノテーションされたテキストは，XHTML へ変換したのちに，指定された XSLT スタイルシートにより，ひまわり XML 形式に変換する。標準で添付しているスタイルシートは，青空文庫の XHTML ファイルから書誌情報などを取得できるようになっている。

XML　XML でアノテーションされたテキストは，指定された XSLT スタイルシートにより，ひまわり XML 形式に変換する。

2.3 アノテーション支援

2.3.1 概要

ここで言うアノテーション支援とは，コーパス構築後に行う，追加的なアノテーションの支援である。そのため，コーパス自体には変更を加えず，stand off でアノテーションするよう設計した。アノテーション結果は，コーパス中の位置情報とともにリレーショナルデータベースに格納される。なお，リレーショナルデータベースエンジン自体も Java 言語で記述されており[1]，拡張された『ひまわり』に標準で同梱している。したがって，構築したコーパスとアノテーション結果を再配布することも容易である。

[1]オープンソースのリレーショナルデータベースエンジン H2 を利用している。

想定するアノテーションとして，(1) コーパスを形態素解析システムや構文解析システムで解析した結果を，コーパスに追加的に自動アノテーションする，(2) 検索結果に注記を人手アノテーションする，ことを考慮した。以下の節では，それぞれの支援方法について，詳しく説明する。

2.3.2 自動アノテーション

タグづけされたテキストに対して，形態素解析や構文解析を行うには，タグを取り除くなどの前処理を行う必要がある。また，解析後も，元々タグづけされていたアノテーションと解析した結果とを統一して検索できるようにする必要がある。

そこで，自動アノテーションの支援では，このような処理を自動的に行うようにする。現時点では，形態素解析システム Mecab と JUMAN の解析結果を扱えるようになっている。データベースに登録する情報は，コーパス中の当該文字列をマークアップし，形態素解析結果に含まれる品詞，活用などの情報を属性として付与するのと同等の情報である。『ひまわり』は，それらの属性をコーパス中の XML のタグの属性と同様に検索することができる。

2.3.3 人手アノテーション

コーパスを利用する際，検索の結果に注記をつけておきたい場合がある。例えば，用例を分類するような場合である。図 2 は,「掘り出す」の用例を語義分けしている例である。

利用者は,「掘り出す」の用例を検索し，それぞれの用例の「メモ 1 」「メモ 2 」欄に注記を加える。「保存」ボタンを押すタイミングで，注記がリレーショナルデータベースに記録される。

注記の形式は 2 種類あり,「メモ 1 」欄が自由記述,「メモ 2 」欄が選択記述となっている。注記欄は追加することも可能である。また，選択記述の項目は，利用者が定義できる。自由記述欄では，選択範囲に値を一括指定するなど，効率的な入力ができるようになっている。これにより，検索した内容をまとめてデータベースへ記録しておく，などの応用が可能である。

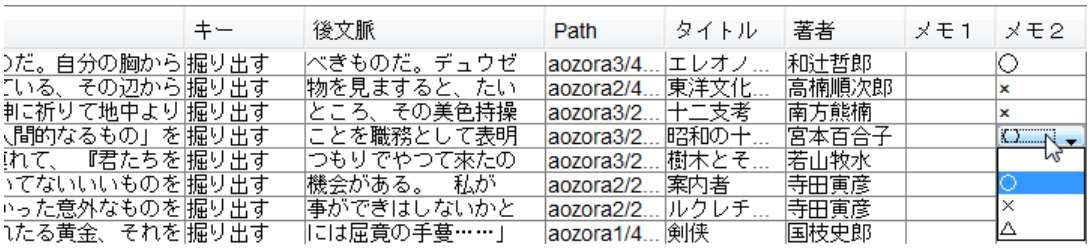

図 2: 人手アノテーションの例

3 実行結果例

以上の支援方法を実装し，拡張版の『ひまわり』として一般に公開している[2]。また，支援機能を利用した例として，大量の「青空文庫」作品を統合し，ひまわり XML 形式に変換した結果 (「青空文庫パッケージ」) も公開している。ここでは，提案した支援手法の実行例として,「青空文庫パッケージ」の構築方法と,「青空文庫パッケージ」への形態素解析結果のアノテーションについて紹介する。

3.1 「青空文庫パッケージ」の構築

「青空文庫パッケージ」は，コーパス構築支援機能を用いて,「青空文庫」で公開されている 10667 作品をひまわり XML 形式のデータに変換したものである。手順は，次のとおりである。

[2] α 版（ver.1.5a02）という位置づけで，2012-10-11 に公開した。なお，このバージョンでは，文字コードポイントの正規化部分は，まだ含まれていない。

(1) 収録対象の作品を選択する。今回は，青空文庫のサイトで公開されている「作家別作品一覧拡充版」から，次の条件に合致する作品を選択した。なお，底本が複数ある作品は,「文字遣い種別」が新字，新仮名の作品を優先している。

- 著作権が切れていること
- XHTML 版が存在すること
- 『ひまわり』用にインポートに成功すること

(2) 「作家別作品一覧拡充版」には，個々の作品の URL が記載されている。その情報をもとに，Web ページのダウンロードツール wget で XHTML 版のファイルを一括ダウンロードした。

(3) ダウンロードしたファイルをコーパス構築支援機能を用いて，ひまわり XML 形式に変換した。ただし，全作品を一括して変換すると，メモリ不足の問題が発生するため，三つに分割し，検索時にそれらを一括検索するようにしている。

3.2 「青空文庫パッケージ」へのアノテーション

「青空文庫パッケージ」に対して，形態素解析結果のアノテーションを行った。使用した形態素解析システムは，MeCab(ver.0.994, IPADIC) である。アノテーションに要した時間は，Ubuntu 12.04 上 (CPU: Intel Xeon E5520 2.27GHz, Memory: 8GB) で約 15 時間かかった。解析結果の形態素数は，85325922 であった。体系的な検索速度の計測は行なっていないが，基本形「投げる」を含む用例を検索したところ，約 23 秒で 4130 例を得ることができた。

4 おわりに

本稿では，個人の研究者がコーパスを構築し，アノテーションするのを支援することを目的として，文字・アノテーションの正規化，自動・手動アノテーションに関する支援方法を提案した。また，全文検索システム『ひまわり』を機能拡張することにより，提案手法を実現した。

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

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

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

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

Mark Davis, Ken Whistler Eds. (2012) *Unicode Normalization Forms, Unicode Technical Reports UAX 15*, http://unicode.org/reports/tr15/

関連 URL

『ひまわり』『えだまめ』 http://www2.ninjal.ac.jp/lrc

『中納言』 https://chunagon.ninjal.ac.jp/

『茶器』http://sourceforge.jp/projects/chaki/

KHCorder http://khc.sourceforge.net/

H2 http://www.h2database.com/html/main.html

Mecab http://mecab.googlecode.com/svn/trunk/mecab/doc/

JUMAN http://nlp.ist.i.kyoto-u.ac.jp/index.phpJUMAN

wget http://www.gnu.org/software/wget/