# 複数の作業者を考慮した形態素データベース修正ツール

山口 昌也
独立行政法人 国立国語研究所
masaya@kokken.go.jp

小島 丈幸
管理工学研究所

## 1. はじめに

近年，コーパスに基づいた形態素解析手法が発達するとともに，タグ付きコーパスの整備が進んでいる。これに伴い，学習用コーパスの構築を支援したり，構築したコーパスを管理・修正するためのツールの重要性が増大している。

そこで，本稿では，高精度の学習用の形態素解析データを構築することを主たる目的として，関係データベースに格納された形態素解析データに対する人手修正ツールを提案する。ここでは,『日本語話し言葉コーパス』[1] の構築における形態素データの人手修正への適用結果について報告する。

形態素解析結果を人手修正するシステムとしては，すでに，タグ付きコーパスの作成支援を行うシステム (VisualMorphs [2] など) やタグ付きコーパスを格納・検索するツール (「茶器」[3] など) が提案され，人手修正のための作業環境が整ってきている。しかし，構築対象のコーパスの規模が大きくなると，(1) 複数の修正者を前提とし，形態素データ全体の整合性を取る仕組みが必要になる，(2) 修正者は必ずしも計算機の専門家であるとは限らないため，修正対象の形態素データを容易に検索できることが必要になる，という問題がある。

これらの問題に対して，本稿で提案するツールは，(1) 関係データベースによる排他制御と最終修正時刻のタイムスタンプを用いることにより，形態素データベース全体の整合性を保持する。また，多数の用例を比較できる KWIC 表示機能を用意することにより，言語的な整合性の保持を支援する。一方，(2) に対しては，GUI によるデータベースへの問い合わせを実現することで，検索の簡便性を確保する。さらに，複雑な問い合わせを必要とする場合については，管理者が修正者に修正対象をまとめて渡せる仕組みを用意する。

この後の節では，まず，2 節で設計時の前提事項について説明した後，3 節で，修正ツールを含めた人手修正環境のシステムの構成を概説する。さらに，4 節で修正ツールの機能を解説し，5 節で提案する修正ツールの適用事例を示す。そして，6 節でまとめを行う。

## 2. 前提

まず，本修正ツールの設計に際して，前提条件となる事柄について述べる。

本修正ツールは,『日本語話し言葉コーパス』(CSJ) の形態素解析データの人手修正用に設計した。CSJ は，主に講演などのモノローグを収録したコーパスであり，音声データの他に，音声データを書き起こした転記テキストが含まれる。本修正ツールは，この転記テキストに対する形態素情報の人手修正に適用する。

人手修正の内訳は，学習用の人手形態素解析における人手修正と，学習結果に基づいた自動形態素解析結果の人手修正からなる。データ量は，それぞれ 100 万語，600 万語と事前に想定した。形態素解析結果の単位は，短単位 [1] とする。

学習用の人手形態素解析における人手修正の手順は，次に示すとおり，網羅的に人手修正を行うものである。

**(1)** 転記テキストを「茶筌」[1)]で形態素解析する。その結果を短単位の品詞体系に変換した上で，関係データベースに登録する [2)]。

**(2)** 本修正ツールにより，講演ごとに人手修正する。全転記テキストを人手修正した後，全形態素データに対して，同一形態素ごとに，整合性が取れているかチェックする。

一方，学習結果に基づいた自動形態素解析結果に対する人手修正の目的は，主として，未知語を発見し，自動形態素解析の精度を向上させることである。したがって，人手修正は，対象を限定して行う。限定方法は，まず，学習用の形態素解析データを内元らの形態素解析システム [4] で学習し，自動形態素解析対象の転記テキストを形態素解析する。そして，各形態素に対して推定される尤もらしさ (確率値) が低い形態素を人手修正する。ここで得られた未知語を辞書登録して，再度形態素解析を行うことにより自動形態素解析の精度を向上させる。

---

1) http://chasen.aist-nara.ac.jp/hiki/ChaSen/

2) 茶筌による形態素解析や短単位品詞体系への変換は，全転記テキストに対して一括して行うのではなく，未知語の辞書登録など，人手修正済みの結果を反映させつつ実行した。

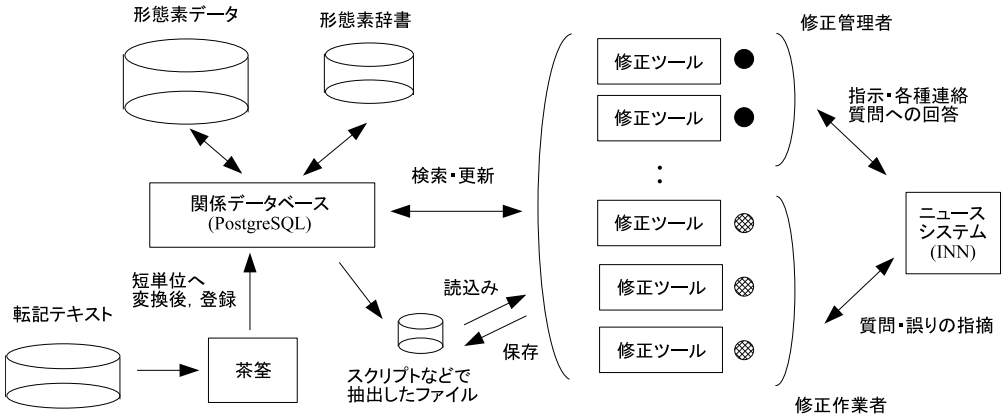


**図 1**: システム構成

## 3. システム構成

本修正ツールを含めた人手修正環境のシステム構成を図 1 に示す。

図 1 のとおり，修正ツールは，形態素データベースのクライアントとして動作する。修正者は，データベースから形態素データの検索・読み込みを行い，修正結果をもとにデータベースを更新する。データベースからの読み込みの他に，スクリプトなどで修正対象の形態素を抽出し，ファイルを介して，修正ツールに読み込むことも可能である。なお，修正ツール自体は，表計算ソフトウェア Microsoft Excel のマクロとして実現している。

形態素データは，関係データベースに格納されており，修正ツールは ODBC を介して，関係データベース (PostgreSQL ver.7.2[3]) にアクセスする。関係データベースには，形態素データの他に，形態素辞書のデータが格納され，修正内容の整合性チェック，および，形態素情報の入力補助のために利用される。

修正者は，全員，言語学の素養があるものとし，同時修正者数は，最大 10 名程度とする。修正者のうち，2 名程度を修正管理者とする。修正管理者は，修正対象の指示や形態素の認定，および，形態素辞書への登録など，コーパス全体の整合性を維持する役割を果たす。

ニュースシステム (INN ver.2[4]) は，複数の修正者が存在することを考慮して，修正上の問題とその解決方法の共有を図るために用意した。具体的な利用方法としては，修正管理者からの連絡・指示や，一般の修正者 (修正作業者) からの質問[5]，それに対する回答などである。

[3] http://www.postgresql.org/
[4] http://www.isc.org/index.pl?/sw/inn/
[5] 形態素分割位置に対する質問や転記テキストの誤りの指摘など

## 4. 修正ツールの機能

### 4.1 概略

すでに述べたように，本修正ツールは，修正対象の形態素データを表計算ソフトウェアに読み込み，その上で修正を行う。形態素データを読み込んだ状態を図 2 に示す。表計算ソフトウェア中の 1 行が 1 形態素に対応する。各列の内容は，左から順に次のとおりである。

- 講演 ID，転記情報
- KWIC(前文脈，出現形，後文脈)
- 代表形 (国語辞典の見出しに相当)，代表表記 (代表形を漢字，仮名などで表記したもの)
- 発音形
- 品詞などの情報 (品詞，活用型，活用形，品詞の下位分類)
- 管理情報 (修正者名，最終更新時間，備考欄，レコード ID，後続する形態素 のレコード ID)

### 4.2 形態素データベースの排他制御

図 1 に示したように，修正ツールは，複数の修正者が同時に作業を行うことを想定している。そのため，データベースに不整合が発生しないように，形態素データベースの排他制御を行う。本修正ツールは，CVS[6] などのバージョン管理システムと同様に,「コピー/変更/更新」型の排他制御を採用した。具体的な排他制御の手順は，次のとおりである。

**(1)** 修正ツールへ形態素データの読み込みを行う。この際，読み込み時だけロックをかける (読み込み終了時にロックを解除する)。修正ツールに読み込まれる形態素データは，形態素データベース上のデータのコピーとなる。

**(2)** 読み込んだ形態素データを修正する。

[6] https://www.cvshome.org/

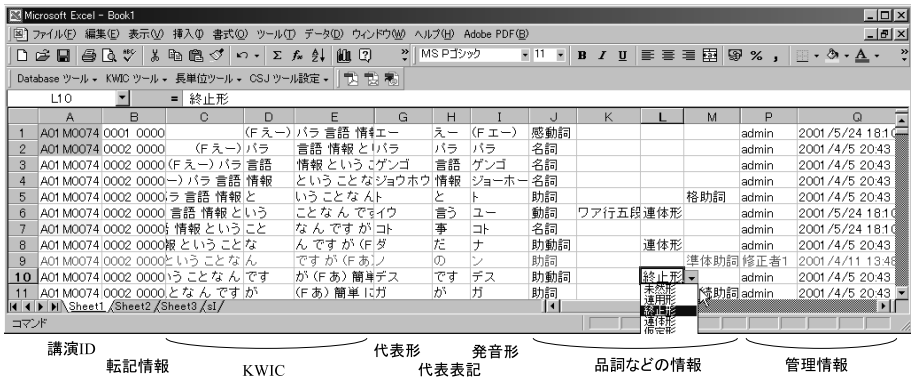


**図 2**: 修正ツール

**(3)** 修正が終了したら，更新処理を行う。まず，(a) 更新対象の形態素データが形態素辞書 (図 1) に登録されているか，(b) 更新対象の形態素データに関して，修正ツールに読み込まれている形態素データの最終更新時刻と形態素データベース側の最終更新時刻とが一致するか，検査する。検査に適合した場合，更新対象のレコードの最終更新時間と修正者名を変更し，ロックをかけつつ，更新する。適合しなかった場合は，修正者に警告した後，更新を中止する。

### 4.3 形態素データの読込み

修正対象の形態素データの読み込み方法には，(1) データベースに対する検索，(2) ファイルからの読み込み，の 2 通りの方法がある。

まず，方法 1 は，図 3 に示した GUI を介して検索した結果を修正ツールに読み込む方法である。GUI では，4.1 節で示したフィールドすべてに対して検索条件を指定できる。生成される検索式は，各フィールドに対する制約値の AND 条件となる。これに付け加えて，前後の形態素に対しても同様の制約を付けて検索することができる。なお，GUI から生成できない検索式については，直接 SQL 文を記述して，問い合わせを実行することも可能である。


**図 3**: 検索用 GUI

また，読み込まれた形態素データをキーとして，キー前後の一定範囲の形態素を読み込むことができる [7]。この検索機能は，特定の形態素をキーとして検索した後，周囲の形態素の情報を修正する必要が生じた場合などに有効である。

一方，方法 2 は，大きく分けて二つの状況で利用する。一つは，SQL 文では目的の修正対象を検索することが困難で，スクリプトなどで抽出する必要がある場合である。もう一つは，方法 1 で読み込んだデータをファイルに保存して，後で編集する状況である。後者は，4.2 節で示した排他制御方式により実現される。この方法は，複数の修正者を想定した修正方式において有用な機能であり，例えば，次の利点がある。

- 修正管理者が，ファイル単位で修正対象の形態素データを修正作業者に割り振ることができるため，作業管理が容易になる (特に，修正対象の形態素データが多い場合)。
- ファイルに作業途中のデータを保存することができるため，修正者自身が進行状況を管理することができる。

### 4.4 形態素データの修正

修正ツールの基本的な修正機能は，形態素の分割位置の修正と，品詞などの付与情報修正である。分割位置の修正については，形態素データベースの観点から次の三つの操作に分類できる (下記，例における ‘/’ は，形態素の分割位置を示す)。

**分割：** 一つのレコードを二つに分割
(例： 外国語→外国/語)

**結合：** 二つのレコードを一つに結合
(例： 外国/語→外国語)

**移動：** 二つのレコード間での文字列の移動
(例： 外/国語→外国/語)

[7] 編集中のシートとは異なるシートへ読み込まれる。

修正ツールは，これらの操作に対する専用のコマンドを用意しており，形態素データベース上の不整合の発生や誤修正を防止する。例えば,「分割」を行う場合は，形態素データベースに新たなレコードを追加することになるが，レコード ID や後続する形態素の ID の整合性を取りつつ，自動的に登録する。また，コーパス本文 (転記テキスト) や講演 ID の変更など，修正者が修正することを禁止されている部分については，修正が加えられないようになっている。

### 4.5 形態素データの修正支援

修正ツールは，修正支援機能として，(a) KWIC 表示，(b) 形態素辞書，(c) メニューによる入力補助，などの機能を持っている。(a) は，図 2 にも示したとおり，出現形に対する KWIC を表示する機能であり，品詞，活用形などの人手修正には不可欠の機能である。また，前後文脈をキーとして，同一形態素の用例をソートすれば，付与情報の整合性維持の支援手段となる。(b) は，図 4 の GUI により，形態素辞書を検索し，その結果を修正中の形態素に反映させる機能である。(c) は，活用形など入力文字列が限定されている項目について，メニューから入力できるようにして (図 2 中のプルダウンメニューを参照)，誤入力を防いでいる。

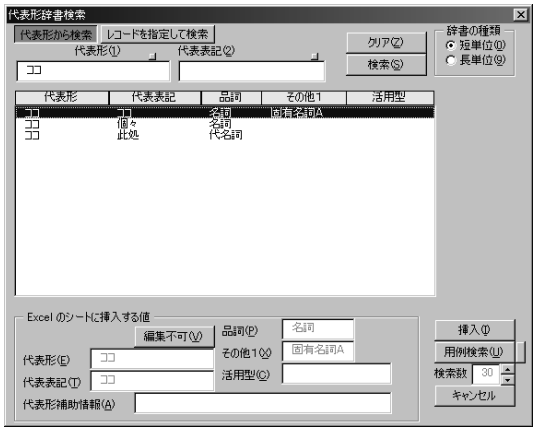


**図 4**: 形態素辞書検索用 GUI

この他にも，表計算ソフトウェア自身の編集機能を使って効率的に修正を行うことができる。例えば，フィルタ機能を利用すれば，修正対象の形態素を編集中に限定することが可能である。また，セルの色やフォントの変更は，修正作業の注釈として利用できるだろう。表計算ソフトウェア自身の編集機能は人手修正に特化した機能ではないが，表計算ソフトウェア自体は一般的に利用されているソフトウェアであるため，人手修正を効率化するだけでなく，修正者が修正ツールを利用する際の障壁を軽減する役割を果たすと考える。

### 4.6 形態素データの更新

修正ツールに読み込まれた形態素データに何らかの修正を加えた場合は，それを形態素データベースに反映する必要がある。修正が加わった行は，文字色が自動的に赤色に変わり，更新対象行となる。更新は，随時，修正者が明示的に実行する。更新処理は，4.2 節 (3) で示した手順で行われる。複数の更新対象行がある場合は，個々の行に対して，順次更新処理が行われる。

## 5. CSJ への適用結果

本修正ツールを CSJ に対する形態素情報の人手修正に適用した結果は，次のとおりである。まず，学習用の人手形態素解析における人手修正については，述べ 1015589 短単位 (異なり 19451) に対してチェックを行い，解析精度は 99.86% (ランダムサンプリング，サンプル数 20000) となった。一方，学習結果に基づいた自動形態素解析結果に対する人手修正では，自動形態素解析結果の約 650 万形態素の中から，新たに 31456 の未知語を抽出し，形態素辞書への登録を行った。

## 6. おわりに

本稿では，形態素データベースに対する人手修正の手法として，複数の作業者を考慮した方法を提案し，『日本語話し言葉コーパス』における形態素情報付与に適用した結果を示した。今後は，さらに大規模なコーパスに対する修正手法や，複数の形式を持ったコーパスを管理・修正する方法について検討する必要があると考える。

**謝辞** 人手修正を行うとともに，さまざまな改良点を提案してくださった修正担当者各氏，自動形態素解析結果を提供してくださった独立行政法人情報通信研究機構の内元清貴氏に心より感謝いたします。

## 参考文献

[1] 前川喜久雄：「日本語話し言葉コーパス」の概要，日本語科学 vol.15，pp.111–133 (2004)

[2] 松田 寛： 品詞タグ付きコーパス作成支援ＧＵＩツール VisualMorphs，情報処理学会研究報告 (2000-NL-137), p.98 (2000)

[3] 松本裕治，高岡一馬ら： タグ付きコーパスの格納／検索ツール「茶器」，言語処理学会第 10 回年次大会発表論文集, p.405–408 (2004)

[4] 内元清貴，野畑 周ら：日本語話言葉コーパスの形態素解析. 言語処理学会第 9 回年次大会発表論文集, pp.113-116 (2003)