書き起こし作業用用字用語辞書の仮名漢字変換システムへの実装と計算機環境の整備

籠宮 隆之[†] 間淵 洋子[††,†] 西川 賢哉[†] 土屋 菜穂子[‡] 小磯 花絵[†]
[†] 国立国語研究所 [††] 東京都立大学大学院 [‡] 青山学院大学大学院

1 はじめに

国立国語研究所・通信総合研究所（現情報通信研究機構）・東京工業大学では，1999 年度〜2003 年度にかけて開放的融合研究「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」[1] の一環として『日本語話し言葉コーパス』[2] を構築してきた．本コーパスは，660 時間以上（約 700 万形態素）の自発音声を格納しており，全ての音声に対して 1) 音声ファイル 2) 書き起こしテキスト 3) 書き起こしテキストの形態素解析結果—— が付与される．また，約 45 時間分の音声に対しては，韻律ラベルや分節音ラベル，係り受け情報などの情報も付与される．

このような大規模なコーパスを効率良く作成するためには計算機の利用が不可欠であり，書き起こし，形態素解析など様々な作業工程で計算機を活用した [3]．本稿では，このうち書き起こし作業を効率良く行なうために作成した仮名漢字変換システム，および書き起こし作業用計算機環境について述べる．

2 書き起こし作業の性格

本コーパスは，音声認識システムで使用する音響モデルならびに言語モデルの作成を主目的の一つとして構築された．言語モデルの構築には，同一の語に対しては単一の表記がされている必要がある．しかし，日本語の表記では 1) 漢字で書くかひらがなで書くか 2) どの漢字を使うか 3) 送り仮名はどうするか —— などの点でゆれが生じやすい．特に，本コーパスのような大規模なコーパスの書き起こし作業は大勢の作業者が長期に渡って行なう必要があったために，どのようにして表記のゆれをなくすかが深刻な課題であった．そこで，本コーパスの表記基準 [4] に従った候補を出力する仮名漢字変換システムを作成し，簡便に表記の統一を行なえるようにした．

また，書き起こし作業は主にネットワークに接続された UNIX ワークステーション上で行なったが，作業者が自宅などで Windows PC を用いて作業することもあった．このため，書き起こし作業のための計算機環境は，複数のプラットホームで同一の操作感を持たせる必要があった．そこで，上記の仮名漢字変換システムを含めた書き起こし作業環境については，さまざまなプラットホームで動作するように設計した．

3 用字用語辞書の仮名漢字変換システムへの実装

3.1 『かんな』形式の選択

計算機上で書き起こしテキストの入力作業を行なう場合に，漢字表記や送り仮名の有無等の表記のゆれを少なくするためには，仮名漢字変換システムで一定の表記にしか変換できないようにするのが良い．そこで，本プロジェクトの書き起こし基準に従って個別の語の具体的な表記を定めた辞書である『用字用語辞書』[5]（約 11 万語）を，NEC が開発した仮名漢字変換システムである『かんな』（`http://www.nec.co.jp/canna/`）の辞書として実装した．

仮名漢字変換システムとして『かんな』を選択したのは，以下のような理由による．

カスタマイズ性 『かんな』はフリーソフトとして配布されており，辞書も含めてユーザが自由に変更できる．また，変換エンジンと文法辞書とが独立しており，文法辞書の変更だけで変換規則が変更できる．

サーバ/クライアントモデル 『かんな』はサーバ/クライアントモデルを採用しており，ネットワーク上に『かんな』サーバを立てておけば，複数のクライアントで辞書を共有できる．

マルチプラットホーム 各種 UNIX 系 OS で広く用いられているのに加え，本プロジェクト開始時（1999 年）には Windows 版の Canna for Windows95 も販売されていた．この Windows 版『かんな』からも『かんな』サーバに接続できるので，ネットワーク上では UNIX と同一の辞書を利用できる．また，Windows 版『かんな』は UNIX 版とバイナリ形式

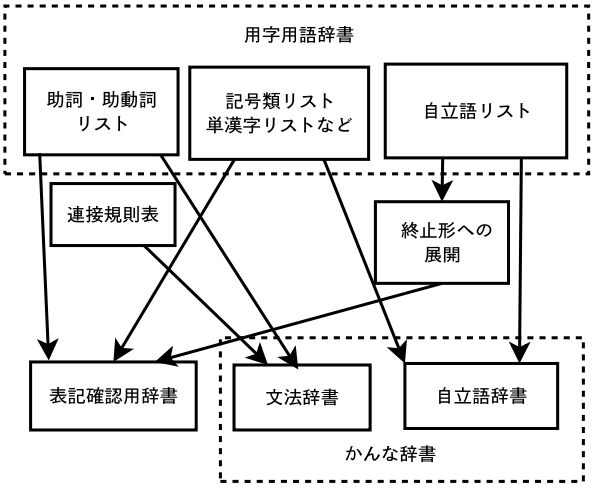

図 1: 用字用語辞書から各種辞書への変換

での辞書の互換性があり，スタンドアロンで UNIX 版の辞書と同一の辞書を使用できる．

用字用語辞書の品詞情報　用字用語辞書は，フリーの仮名漢字変換辞書である Pubdic+をベースにして作成された．Pubdic+は，UNIX で広く用いられてきた『かんな』，Wnn，SJ3 用の辞書として作成されたので,『かんな』の品詞情報も付与されている．用字用語辞書の編集にあたっては，この『かんな』の品詞体系をもとに品詞の分類を行なった．そのため，用字用語辞書からは容易に『かんな』形式へ変換できる．

3.2　仮名漢字変換用辞書の生成

用字用語辞書は，語種・品詞によって分割された，幾つかのリストより構成される．

自立語リスト　メインのリストである．現在約 11 万語が登録されている．用字用語辞書のベースとなった Pubdic+由来の『かんな』の品詞情報が付与されている．

記号類リスト，単漢字リストなど　単漢字や記号類のリストなど，活用の無い語のリストである．品詞情報は付与されていない．

助詞・助動詞リスト　助詞・助動詞などの付属語のリストである.『かんな』の付属語リストを元に口語表現を適切に変換できるようにするなどの拡張を施したものである.『かんな』の品詞情報が付与されている．

用字用語辞書
○　おこな:行な:あわ行 (#W5r)
×　おこな:行:あわ行 (#W5r)〔→「行な (う)」〕

表記確認用辞書			
○	行なう	おこなう	[動詞]
×	行う	おこなう	[動詞]　〔→「行な（う)」〕

表 1: 用字用語辞書から表記確認用辞書への変換

それぞれのリストに付与されている情報を元に,『かんな』の仮名漢字変換辞書形式に変換する（図 1 参照）．

自立語リストからは，用字用語辞書で「使用できる」とされた表記 [5] のみを取り出し,『かんな』の自立語辞書に登録する．記号類や単漢字のリストには品詞情報が付与されていないので，適切な品詞を付与した後に『かんな』の自立語辞書に登録する．また，助詞・助動詞リストおよび連接規則表からは，文法辞書を作成する．連接規則表は，各品詞ごとに許可される連接関係を定義したものである．この連接規則表にも,『かんな』のオリジナルのものに対して口語表現形式の品詞を新たに定義するなどの変更を加えている [5]．

なお，仮名漢字変換辞書を作成する際には，仮名漢字変換辞書と同時に，後述する表記確認用の辞書も用字用語辞書から生成する（図 1 参照）．

3.3　表記確認用辞書の生成

仮名漢字変換システムを整備すれば，使用可能な表記を得ることはできる．しかし，表記の統一を徹底するためには，1) どの表記が使用できないか．2) 使用できない場合にはどのような表記をとれば良いか．——も明示する必要がある．そこで，作業者が書き起こしの途中で逐次参照するための表記確認用辞書も生成した．この辞書には「使用できる表記」だけでなく,「使用できない表記」も明示され，また，書き起こしの際に参照できる豊富な注記情報が付与されている．用字用語辞書には，これらの情報が全て含まれている [5]．

この表記確認用辞書を用字用語辞書から生成するにあたり，書き起こし作業者が平易に読める形式に変換した．用字用語辞書に付与されている品詞情報は『かんな』の品詞情報であるが，この品詞情報はニューメリックな形式で記述されており，一般には馴染みのない形式である．また，自立語リストの見出語は語幹のみが記載されており，やはり一般には馴染のない形式である．そこで，1)『かんな』の品詞情報を一般的な形式に変換する．2) 自立語の語幹を終止形に展開する．——

図 2: 書き起こし環境のスクリーンショット

という変換を行なった（図 1，表 1 参照）.

4 書き起こし作業用計算機環境

これまでに述べた仮名漢字変換システムおよび表記確認用辞書を効率良く利用するための書き起こし作業用統合環境を作成した．書き起こし作業用統合環境に必要とされる機能は，1) 当該箇所の音声を簡便に聴取できること．2) 今回作成した『かんな』用仮名漢字変換辞書を用いて仮名漢字変換ができること．3) 今回作成した表記確認用辞書を効率良く利用できること．—— などである.

この統合環境は，Emacs エディター上に構築した．Emacs エディターを用いた理由として，1) Emacs エディターは各種 UNIX，Windows，MacOS など，多くのプラットホームで使用できる．2) メーラーや電子辞書検索クライアントなどのプログラムが Emacs 上で動作するので，これらと連携してより効率良い書き起こし作業が行なえる．3) Emacs のマクロだけで実装された『かんな』のクライアントである YC を用いれば，Canna for Windows95 を用いなくても Windows 上で『かんな』を用いた仮名漢字変換が可能．—— などが挙げられる.

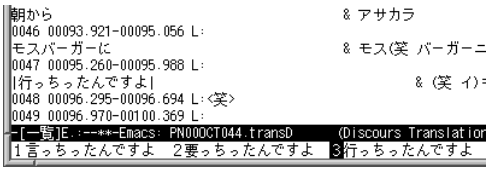

図 3: 口語表現の変換

図 4: 選択した語で表記確認用辞書を検索

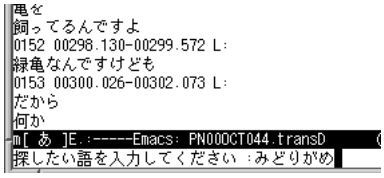

図 5: 検索したい語を入力して表記確認用辞書を検索

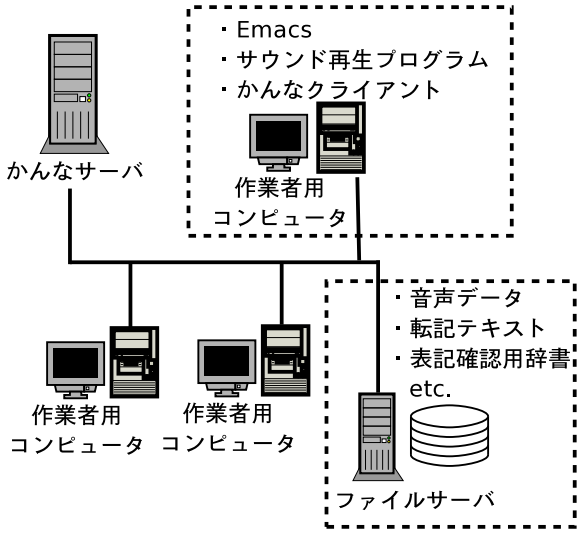

図 6: ネットワークで接続された作業環境

図 2 に，書き起こし用統合環境を用いて書き起こし作業を行っている様子を示す．メインウィンドウでは，「ふれあいひろばというのが」を仮名漢字変換している．『かんな』の変換方法は連文節変換形式であるため，入力はそれぞれ「ふれあい」「ひろばと」「いうのが」という文節に区切られる．図 2 のスクリーンショットでは「ふれあい」の変換候補を提示しているが，用字用語辞書に登録されている表記が「触れ合い」のみであるので，漢字仮名交じりの変換候補は「触れ合い」のみが提示されている．ただし,『かんな』は常に入力された仮名文字列に対する平仮名表記とカタカナ表記も提示するので,「ふれあい」「フレアイ」も提示される．作業者には，この平仮名表記とカタカナ表記の候補は変換候補として利用しないように指示した．

図 3 には，口語表現である「いっちったんですよ」を変換している様子を示す．口語表現で頻出する「〜ちった」を仮名漢字変換システムの文法辞書に組み込んでいるため,「行っちったんですよ」などが変換候補として提示される．

図 2 の左上のポップアップウィンドウでは,「ふれあい」はどのような表記を用いれば良いかを，表記確認用辞書を検索して調べている．表記確認用辞書の検索には，1) メインウィンドウで検索したい語を選択する（図 4）．2) ミニバッファに検索したい語を入力する（図 5）．—— のいずれかの操作方法をとる．これらの検索は，メインウィンドウ上のキー操作により簡便に行なえる．

ネットワークに接続された環境では，作業者の用いるクライアントマシンはネットワーク上のかんなサーバに接続して仮名漢字変換を行なう．また，表記確認用辞書もファイルサーバ上のものを共有する（図 6 参照）．これにより複数の作業者で同一の辞書を使用でき，異なるバージョンの辞書を使用してしまう，などのエラーを防ぐことができる．

音声ファイルの再生も，書き起こしテキスト中の当該箇所でのキー操作により行なう．なお，音声ファイルを再生するのには外部コマンドを用いる．外部コマンドは 1) ファイル名 2) 再生開始位置 2) 再生終了位置 —— を指定すれば当該箇所の音声を再生するプログラムを使用する．UNIX 上では商用の音声分析用ソフトウェアである ESPS/XWaves+付属のコマンドを用いたが，Windows 用には適当なものがなかったので，別途 Playwav（`http://www2.kokken.go.jp/~kagomiya/playwav.html`）を作成した．

5 おわりに

以上,『日本語話し言葉コーパス』の書き起こしテキストを効率よく作成するための仮名漢字変換辞書および書き起こし統合環境について述べた．これらは近日中に公開する予定である．公開方法等については，後日国語研のウェブサイト等でアナウンスする予定である．

参 考 文 献

1) 古井貞熙, 前川喜久雄, 井佐原均：科学技術振興調整費開放的融合研究推進制度 — 大規模コーパスに基づく『話し言葉工学』の構築 —, 日本音響学会誌, Vol. 56, No. 11 (2000).

2) 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, (2004).

3) 前川喜久雄, 菊池英明, 籠宮隆之, 山口昌也, 小磯花絵, 小椋秀樹:『日本語話し言葉コーパス』構築における計算機利用, 日本語学, Vol. 20, No. 13 (2001).

4) 小磯花絵, 土屋菜穂子, 間淵洋子, 斉藤美紀, 籠宮隆之, 菊池英明, 前川喜久雄:「日本語話し言葉コーパス」における書き起こしの方法とその基準について, 日本語科学, Vol. 9, (2001).

5) 間淵洋子, 西川賢哉, 土屋菜穂子, 相馬さつき, 籠宮隆之, 小磯花絵, 前川喜久雄:『日本語話し言葉コーパス』書き起こしの為の用字用語辞書の作成, 本号所収 (2005).