「近代文語 UniDic」「中古和文 UniDic」を利用した総索引作成システムの開発

小木曽 智信　　須永 哲矢
人間文化研究機構 国立国語研究所

日本語史の研究において検索のための総索引（concordance）はたいへん重要な位置を占めている。昨今，歴史的資料を対象とした形態素解析が実用化されたことにより，総索引を半自動的に作成することが可能になった。しかし，一般的な日本語研究者にとって既存のソフトウェアを用いて総索引を作成することは難しかった。そこで，形態素解析結果から容易に総索引を作成することのできるソフトウェアの開発を行った。

Development of the publishing system of Japanese concordances using morphological analysis with *Kindai-Bungo UniDic* and *Chûko-Wabun UniDic*

Toshinobu Ogiso　　Tetsuya Sunaga
National Institutes for the Humanities,
National Institute for Japanese Language and Linguistics

Concordances of text materials occupy a very important position in research of the history of the Japanese language. Recently, computerized morphological analysis for historical Japanese texts has been made practical, and it has become possible to create concordances semi-automatically. However, it has been difficult for general Japanese linguists to make concordances with existing software until now. Accordingly, we developed software which makes it possible to create concordances easily.

1. はじめに

日本語・日本文学の研究，なかでも日本語の歴史的な研究において，用例検索のための総索引（concordance）はたいへん重要な位置を占めている。昨今，歴史的資料を対象とした形態素解析が実用化されたことにより，かつては膨大な手作業を必要とした総索引の作成を半自動的に行うことが可能になった．しかし，一般的な日本語研究者にとって既存のソフトウェアを用いて総索引を作成することは容易ではなかった。そこで，形態素解析結果から容易に総索引を作成することのできるソフトウェアの開発を行った。

本発表では，歴史的資料を対象とした形態素解析辞書「近代文語 UniDic」「中古和文 UniDic」を利用して，紙ベースの総索引を作成するためのシステムを提案する。対象資料のテキストファイルから，形態素解析（及び必要に応じた人手修正）を経て，文脈付き総索引の生成を可能にするものである。

2. 日本語学と総索引

語の用例を見つけ出し，それに基づいて議論を行うことが研究の出発点となる日本語学において，語の検索のための総索引はたいへん重要である。特に文法性判断などで内省がきかない歴史的研究において，用例の実態調査は研究の基礎となる重要な役割を果たすものである。そのため，多くの文学作品や国語資料について総索引が作られ，研究に利用されてきた。国文学研究においても総索引は研究に欠くことのできないものとして広く用いられている。

今日では主立った文学作品の総索引は出そろった感があるが，比較的マイナーな作品を対象としたものや，別系統の伝本にもとづくものなどはいまだ不十分であり，現在でも新たな総索引が作成・刊行されている。資料の電子化・データベース化が進む中でも，これまでと同じ使い勝手を求める声は少なくなく，紙ベースの書籍形態の総索引の需要は大きいようである。近年における日本語研究分野の科研費採択課題から見ても，総索引を新たに作ろうとする試みが続けられていることがわかる。

初期に作成された総索引の多くは，文脈なしの自立語索引であったが，日本語史研究においては付属語も重要な調査対象である。そのため，今日では付属語までを含み，KWIC 形式の文脈を付与するものが一般化している。

このような総索引を作成するに当たってはコンピュータを用いた何らかの手助けが必要とされる。文脈生成や見出し語のソート，整形については，レポート出力可能なデータベースソフトなどを用いることで対応できるが，一般的な日本語・日本文学の研究者にとってはハードルが高い作業である。また，単語に分割して品詞や読みなどを付与する作業は，人手によらなければならずたいへんな手間を要する。しかし，形態素解析技術の発達により，その大部分を自動化することが可能になった。

2. 歴史的資料を対象とした形態素解析

現代日本語の形態素解析システムについては，京都大学の JUMAN や奈良先端科学技術大学の ChaSen，それに続く MeCab などの成功により，以前から実用的な精度での解析が可能になっていた。しかし，日本語研究用の総索引が対象とするような歴史的資料については，長らく十分な精度での解析を行うことができなかった。

しかし，近年，「近代文語 UniDic」（[1][2][4]）によって明治期の文語論説文の形態素解析が可能になったほか，「中古和文 UniDic」（[3][4]）によって，平安時代の和文形資料の解析も可能になってきた。特に後者が対象とする文体は，後の時代の資料でも用いられており，総索引作りの対象となることの多いものである。

この形態素解析辞書を用いることにより，単語分割や品詞・読みの付与といった，総索引作成の手間を大幅に減らすことが可能になると考えられる。現在，歴史的資料を対象とした形態素解析システムの解析精度は 96%程度であり，そのまま総索引の元データとするには問題が残る。しかし，これに人手によるチェックと修正を加えることで，総索引作成用として十分な精度に高めることが可能である。

UniDic は標準の形態素解析結果として表 1 に示す属性を出力する。これ以外にも，設定により，書字形基本形やアクセント型などの豊富な情報が出力可能である。

表 1 のうち語彙素・語彙素読み・品詞・活用型・活用形・書字形・発音形（または語形）の組により，辞書中の見出し語として一意に同定することができるようになっている。

表 1 UniDic の形態論情報

属性名	説明
語彙素	辞書見出しの代表表記
語彙素読み	辞書見出しの読み（カナ）
語形	異語形を区別する形（カナ）
品詞	品詞（大-中-小分類）
活用型	活用型（活用語のみ）
活用形	活用形（活用語のみ）
書字形	テキストに出現した表記形
発音形	読み上げ用の形（現代読み）
語種	和語・漢語・外来語等の別

総索引の作成に当たっては，このうち「語彙素」「語彙素読み」「語形」「品詞」「活用型」「書字形」「語種」を出力した。また印刷される結果には出力しなかったものの「活用形」を索引のソートに利用した。

3. 総索引作成システム

「近代文語 UniDic」「中古和文 UniDic」を利用して，日本語学の研究者が容易に総索引を作成することができるようにするために，UniDic の出力形式を入力として，最終的に PDF（ないし HTML）形式の総索引を出力とする総索引作成システムを開発した。システムの全体の流れは図 1 のようになる。

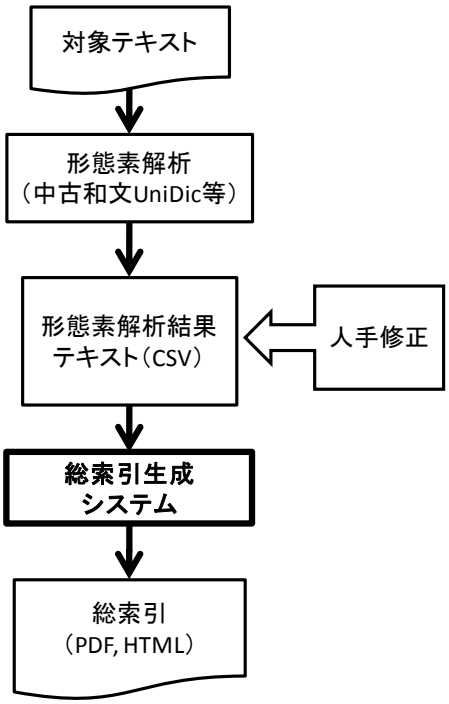

図 1　総索引作成システムの流れ

対象となるテキストは，プレーンテキストまたは XML ファイルに対応する。XML 形式の場合も，タグ付けされた情報は，後述する位置情報以外には用いない。基本的にプレーンテキスト化し，改行と句点を基準にして文を区切り，これを形態素解析への入力とする。

形態素解析そのものは，UniDic 標準の解析方法による。UniDic には形態素解析を手助けするインターフェイスプログラム「茶まめ」が付属しており，コンピュータに不慣れなユーザでも容易に形態素解析を行うことが可能になっている。茶まめを用いた解析結果の出力は表 1 に示した情報が含まれるタブ区切りテキスト形式である。

先述したとおり，形態素解析結果をそのまま総索引に用いることは，精度上問題がある。したがって，多くの場合には人手による修正を行う必要があるが，その際にも，形態素解析結果と同じ表形式を保ったまま作業を行うことは可能である。そのため，本システムは形態素解析結果そのままの入力形式を想定しているが，解析結果をそのまま用いた全自動の総索引生成と，人手修正後のデータを用いた総索引生成との両方に対応している。

形態素解析結果の修正

総索引生成システムの入力である形態素解析結果はタブ区切りテキスト形式であり，必要に応じて表計算ソフトなどを用いて修正することができる。しかし，形態素解析結果だけを見て，テキストの他の箇所との整合性を保ちながら解析結果の修正作業を行うことは必ずしも容易ではない。テキストの形態素解析結果（コーパス）と形態素解析辞書の見出し語をデータベース上で関連づけ，辞書を参照しつつテキスト解析結果の修正を行うことができれば，比較的容易に修正作業を行うことができる。

今回は，この修正作業用のシステムとして国立国語研究所の「形態論情報データベース」[5] [6]を用いた。しかし，これは「現代日本語書き言葉均衡コーパス」の構築のために運用しているデータベースシステムであり，一般に利用可能なものではない。今後，奈良先端科学技術大学院大学で開発されているコーパス検索・管理ツール「茶器」を用いることで，一般ユーザにもコーパス整備が可能な環境を構築していきたいと考えている。

形態論情報の簡略化

UniDic において，たとえば普通名詞は「名詞-普通名詞-一般」といった，大分類-中分類-小分類の形で表される。しかし，一般の総索引利用者にとっては，この長い形式は煩雑である。また，「大分類」や「中分類」だけを取り出しても，今度はおおざっぱになりすぎたり，品詞によっては細かすぎたりといった問題が生ずる。そこで，索引を作成する段階で，この長い品詞を適切に簡略化させる機能を用意した。

UniDic の品詞と簡略品詞の対応表を用意し，これを読み込んでパターンマッチによって置き換えるものである。品詞対応表は単純なテキストファイルであり，索引の作成者が自由にカスタマイズすることが可能になっている。

このような簡略化は，活用型・活用形でも必要となる。UniDic の出力では，活用型は「文語四段-カ行」，活用形は「終止形-一般」などと，これも索引の利用者から見た場合には煩雑な長い形で出力されている。そこで，これらも品詞と同様の対応表によって簡略化を行っている。

本文中の位置情報

索引において，語の出現箇所は元のテキストにかえって本文や注釈などを確認するために極めて重要な情報である。したがって，総索引作成システムでは位置情報を出力できるようにする必要がある。

プレーンテキストを入力とした場合には，すべてのテキストが形態素解析の対象となるため，位置情報をタグ付けして残すことができない。テキストファイルから取得可能な位置情報としては，ファイル先頭からの文字数・文数・語数などが考えられる。文字数は数字が大きくなりすぎ扱いにくく，語数は形態素解析結果の修正によって変化してしまうため，今回は文数を採用することとした。

XML ファイルの場合には位置情報をタグ付けしておくことができる。そこで，総索引作成システムのオプションで形態素解析対象となった XML ファイルを指定することにより，位置情報を索引に埋め込むことができるようにした。タグの形式としては，

<info position=”位置情報”/>

という簡単な空要素タグだけを参照して利用することとした。位置情報は，終了位置を表すものとし，索引には，おのおのの語から

見て後方に最初に現れる位置情報を出力している。位置情報と語との関連づけは，ファイル中の文字位置によっている。

なお，位置情報については，ページ区切り，行区切りなどを階層化して入力することも考えられるが，現在のところ，簡潔さを優先して一つのタグを出力するのみとなっている。

索引の出力

総索引生成システムは，整形した文脈付き総索引を出力する。文脈は，形態素解析結果の書字形（表層の出現文字列）を出現順に組み上げて生成する。見出し語は，読み（UniDic の「語彙素読み」）を基本に代表表記（同「語彙素」）・品詞・活用型をキーとして並び替え・グループ化し，アイウエオ順に出力する。

出力形式は，オプションにより，TeX（PDF）形式と HTML 形式に対応する。TeX（PDF）形式は高品質な組み版・印刷に供するためのものであるが，TeX ファイルをコンパイルして PDF を出力するために Unicode 対応の TeX（upLaTeX）環境が必要となる。一方，HTML 形式は特別な環境なしに容易に利用可能にするためのもので，CSS による簡易なデザインを施している。CSS の修正により，デザインを変更することも可能となっている。索引のサンプルは，末尾の【付録】を参照されたい。

4. 総索引作成例－『恋路ゆかしき大将』文脈つき総索引

このシステムを利用して，実際に索引サンプルを作成した。対象とした作品は，鎌倉期の擬古物語の一つ，『恋路ゆかしき大将』[7]（作者不詳，宮田光校訂・訳注）である。この作品を含め，中世の物語作品は，総索引が作られていないものが多く，本システムの活用が期待されるジャンルのひとつであるといえる（表 2）。

『恋路ゆかしき大将』総索引の作成は，本文を OCR によってテキストデータ化することから始めた。テキストデータ化の対象ページ数は計 117 ページで，二段組みの上段に本文，下段に現代語訳があるうちの上段のみをテキスト化した。文字数約 7.5 万，短単位語数 4.5 万（記号・句読点含む）。

作業手順とそれに要した日数は表 3（次頁）に示したとおりである。作業者 1 名，1 日あたりの作業時間はおおむね 6 時間程度である。

表 2　笠間書院『鎌倉時代物語集成』所収作品の索引の有無

作品名	索引の有無
あきぎり	
あさぢが露	自立語索引あり
あまのかるも	総索引あり
在明の別	
石清水物語	総索引あり
いはでしのぶ	
風につれなき物語	
風に紅葉	語句索引あり
苔の衣	
木幡の時雨	総索引あり
恋路ゆかしき大将	
小夜衣	総索引あり
雫に濁る	
しのびね物語	
白露	
住吉物語	総索引あり
とりかへばや	総索引あり
兵部卿物語	
松陰中納言物語	
松浦宮物語	総索引あり
むぐらの宿	自立語索引あり
無名草子	総索引あり
八重葎	
(別本)八重葎	
山路の露	総索引あり
夢の通ひ路物語	
夜寝覚物語	総索引あり
我身にたどる姫君	
雲隠六帖	
下燃物語	
豊明絵草子	
なよ竹物語	総索引あり
掃墨物語	
葉月物語	

今回の総索引作成にあたっては，書籍形態の本文しか存在しない状態から開始したが，それでも形態素解析および索引生成の自動化により，23 日で総索引が完成した。実際には本文の OCR 結果に対する校正をどの程度行うか，形態素解析結果にどの程度細かく修正を施すかなどによって作業時間は変わってくるが，およそ 3 週間で，4.5 万語程度の総索引

が作成できたことになる。なお，今回は本文校正，形態素解析結果ともに，十分な確認作業を行っている。

表 3　総索引作成の作業手順と作業日数

作業手順		作業日数
1	書籍をスキャン	1 日
2	OCR による本文のテキストデータ化と校正作業	3 日
3	タグ付け，XML データ化	3 日
4	「中古和文 UniDic」による形態素解析	（自動）
5	形態素解析結果人手修正	16 日
6	総索引生成システム	（自動）
（計）		23 日

形態素解析には「中古和文 UniDic」を用いた。この解析辞書は，主として平安時代の和文資料を対象としたものであるが，鎌倉期以降の作品であっても，中古和文に類する文体の資料に対しては，ほぼ同程度の解析精度が期待できる。実際の作業においても，平安時代作品を解析した場合に比して不都合は感じられなかった。

5. おわりに

歴史的資料を対象とした形態素解析辞書により，古典テキストの高度な利用が可能になった。しかし，日本語・日本文学の一般的な研究者にとって，形態素解析やデータベースの操作といった技術はいまだに敷居の高いものであって，なかなか利用が進んでいない。本システムのように，新しい技術による成果を従来からある研究用の資源の形式で提供することによって，技術への理解とその普及を図るとともに，新たな研究用資源の構築につなげていくことができればと考えている。

参考文献

[1] 小木曽智信・小椋秀樹・近藤明日子（2008）「近代文語文を対象とした形態素解析辞書・近代文語 UniDic」『言語処理学会第 14 回年次大会予稿集』pp.225-228

[2] 小木曽智信（2009）『近代文語文を対象とした形態素解析のための電子化辞書の作成とその活用』国立国語研究所・科研費報告書 19720110

[3] 小木曽智信・小椋秀樹・田中牧郎・近藤明日子・伝 康晴（2010）「中古和文を対象とした形態素解析辞書の開発」情報処理学会研究報告 人文科学とコンピュータ Vol.2010-CH-85(No.4) pp.1-8

[4] 「近代文語 UniDic」「中古和文 UniDic」http://www.kokken.go.jp/lrc/index.php?UniDic

[5] 小木曽智信・中村壮範（2009）『『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装』国立国語研究所内部報告書

[6] 「茶器」http://chasen.naist.jp/hiki/ChaKi/

[7] 宮田光・稲賀敬二（2004）『中世王朝物語集 8「恋路ゆかしき大将」「山路の露」』笠間書院

【付録】『恋路ゆかしき大将』文脈つき総索引サンプル

HTML 版

アウ 【合う】〔和〕 アウ （動詞・四段-ハ行） 10

1-35-1	ならず。この君達の母君はとく失せ給ひて、うち	合は	ぬ事多く、あはれなる御さまなるに、大臣御心落
2-62-2	かりける大臣の御覚えのやうを、けしからずうち	合は	ず思さるれど、例の、ただわが御心もなきさまに
2-65-18	けれど、世に旧り、ただ人の母君にて、またうち	合は	ぬ事にもどきあらんを憚り思されて、滞りしかど
3-122-13	ましたり。かしこには、例の御物怖ぢに、御目も	合は	ずみじろき臥させ給へるまだ暁、二品宮より、雪
5-152-17	、かく品遅れてもてなし給ふ事、日ごろも御心に	合は	ず思ひたてまつり給ふ大宮の御事なれば、うち合
5-152-18	はず思ひたてまつり給ふ大宮の御事なれば、うち	合は	ず思したり。いみじうのたまひ慰めつつ、顧みが
5-187-16	されたり。大宮は、またそれにつけて、いよいよ	合は	ぬ世にますがみおはしませば、いづ方も苦しく、
5-200-7	る方はおはしまさぬ御本性に、いとあいなし、事	合は	ぬ事、とけしき御覧じたるべし。女君はいろいろ
5-205-17	大将は心も空に、「あぢきなのさまや」と、目も	合は	ねど、いづ方とて疎かなるべきならねば、出で給
1-26-8	もあやまたるるほどなるを、御供の人どもも興じ	あへ	り。後夜をばつかで、晨朝をほかより夜深くつく

アエカ 【あえか】〔和〕 アエカ （形状詞） 1

3-115-7	高うよしあるさま、あさましきまで驚かれ給ふ。	あえか	になつかしうなどはあらねども、際あがりてめづ

アエナイ 【敢え無い】〔和〕 アエナシ （形容詞・ク） 1

5-214-2	身には何事もかひなかりけりと思ひ知らるれば、	あへなし	。この御さまの心苦しさに添ひおはしつつ、いと

アエモノ 【肖物】〔和〕 アエモノ （普通名詞） 1

1-12-16	。御心ざしいともあらまほしきさまにて、世の人	あえ物	にも聞こえさせしを、思し喜びて父大臣は失せ給

アエル 【零える】〔和〕 アユ （動詞・下二段-ヤ行） 1

5-179-3	けり。何の行方も知らぬ若き女の卑しからぬが乳	あゆる	に抱かせたてまつりて、ただ一人乗せ給へりけり

アオイ 【葵】〔和〕 アオイ （普通名詞） 1

5-153-5	、命ながさも怨めしく、つらき事言はん方なし。	葵	かざして見たてまつり初めしより、今日までの御

アオイ 【青い】〔和〕 アオシ （形容詞・ク） 1

1-27-11	寄りてぞ言ふなる。ほども経ず、濃き蘇芳の袒に	青き	一重着て、大きらかなる童の、薄様の端に書きた

アオグ 【仰ぐ】〔和〕 アオグ （動詞・四段-ガ行） 2

3-113-13	りけるに侍るを、おのれも、思ひの外に今は君と	仰ぎ	たてまつり侍れば、申し出づる事のついでもやと

TeX(PDF） 版

アウ 【合う】 〔和〕 アウ （動詞・四段-ハ行） 10 例

1-35-1 ならず。この君達の母君はとく失せ給ひて、うち【合は】ぬ事多く、あはれなる御さまなるに、大臣御心落
2-62-2 かりける大臣の御覚えのやうを、けしからずうち【合は】ず思さるれど、例の、ただわが御心もなきさまに
2-65-18 けれど、世に旧り、ただ人の母君にて、またうち【合は】ぬ事にもどきあらんを憚り思されて、滞りしかど
3-122-13 ましたり。かしこには、例の御物怖ぢに、御目も【合は】ずみじろき臥させ給へるまだ暁、二品宮より、雪
5-152-17 、かく品遅れてもてなし給ふ事、日ごろも御心に【合は】ず思ひたてまつり給ふ大宮の御事なれば、うち合
5-152-18 はず思ひたてまつり給ふ大宮の御事なれば、うち【合は】ず思したり。いみじうのたまひ慰めつつ、顧みが
5-187-16 されたり。大宮は、またそれにつけて、いよいよ【合は】ぬ世にますがみおはしませば、いづ方も苦しく、
5-200-7 る方はおはしまさぬ御本性に、いとあいなし、事【合は】ぬ事、とけしき御覧じたるべし。女君はいろいろ
5-205-17 大将は心も空に、「あぢきなのさまや」と、目も【合は】ねど、いづ方とて疎かなるべきならねば、出で給
1-26-8 もあやまたるるほどなるを、御供の人どもも興じ【あへ】り。後夜をばつかで、晨朝をほかより夜深くつく

アエカ 【あえか】 〔和〕 アエカ （形状詞） 1 例

3-115-7 高うよしあるさま、あさましきまで驚かれ給ふ。【あえか】になつかしうなどはあらねども、際あがりてめづ

アエナイ 【敢え無い】 〔和〕 アエナシ （形容詞・ク） 1 例

5-214-2 身には何事もかひなかりけりと思ひ知らるれば、【あへなし】。この御さまの心苦しさに添ひおはしつつ、いと

アエモノ 【肖物】 〔和〕 アエモノ （普通名詞） 1 例

1-12-16 。御心ざしいともあらまほしきさまにて、世の人【あえ物】にも聞こえさせしを、思し喜びて父大臣は失せ給

アエル 【零える】 〔和〕 アユ （動詞・下二段-ヤ行） 1 例

5-179-3 けり。何の行方も知らぬ若き女の卑しからぬが乳【あゆる】に抱かせたてまつりて、ただ一人乗せ給へりけり

アオイ 【葵】 〔和〕 アオイ （普通名詞） 1 例

5-153-5 、命ながさも怨めしく、つらき事言はん方なし。【葵】かざして見たてまつり初めしより、今日までの御

アオイ 【青い】 〔和〕 アオシ （形容詞・ク） 1 例