自然言語処理 32－6
（1982. 8. 2）

本文研究の立場からみたテキスト・データベースの機能について

田嶋一夫・星野雅英（国文学研究資料館）

1．はじめに
2．文学研究におけるテキスト情報の活用
3．索引誌作成の現状と問題点
4．テキスト・データベースの必要性

1．はじめに

人文科学という学問において、テキスト（本文）そのものが研究対象となる分野は、文学、歴史学、言語学、国語学、心理学、等が考えられる。なかでもテキストに対する研究比重の高いところは文学である。ここでは、その研究の全体にわたってテキスト分析がかかわっていると言える。その他においても、いわゆる“本”が研究対象となるところでは、そのテキストの緻密な分析が必らず伴うであろう。こうした意味でテキストのデータベースの活用される分野は、人文科学研究のすべてと言っても過言ではない。しかし、もっとも多くのかかわりを持ち、重要な資料としての意味を持つのは文学研究である。

しかし、文学研究の中でも、テキストが分かち書きされている英文学や仏文学等の西欧文学と、日本文学のように分かち書きされていない分野では、大きな差違がある。ここでは日本文学のテキストを念頭におきつつ、テキスト・データベースのあり方について考察する。

2．文学研究におけるテキスト情報の活用

テキストを主な研究対象とする文学研究において、テキストの情報をどのように活用するかを考えると、およそ次の2点になる。

1）本文批判の際の資料として

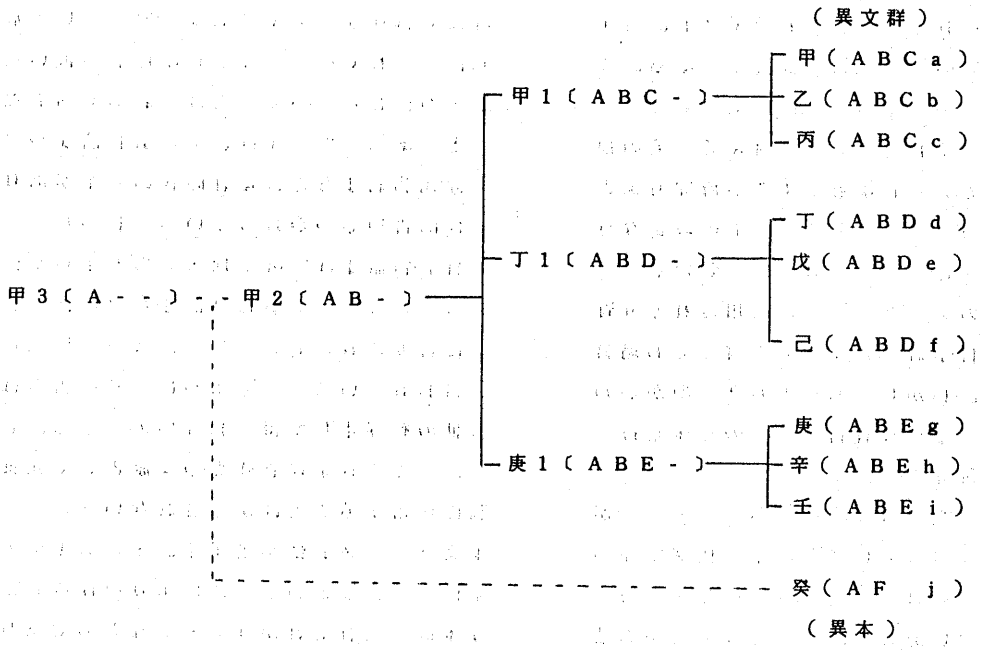
2）作品分析、本文分析の際の資料として

まず、本文批判の問題を考えると、文学研究（古典）のテキストは、写本という形で伝えられている。しかし原作者の書いた原本が伝存していることはきわめて稀有なことである。たとえば、『伊勢物語』や『源氏物語』などは鎌倉時代の写本があり比較的古いものが残っている。しかし、『竹取物語』は室町末期までの写本しか伝わっていない。『蜻蛉日記』などは江戸初期に書写されたものまでしか残っていないと言われている。このように多くの古典は後世の転写本しか得られないのである。したがってこれを成立時点の文献として無批判に使用することはゆるされない。ここに本文批判の必要性が生ずる。またテキストを伝える書写という方法により製作された写本は、それぞれがユニークなものであり、絶対に同一の写本はありえない。しかも各写本は必らずその依拠した親本の特徴的な本文の影（因子）を保持しているはずである、と考えられる。とするならば、現存するばらばらで相互に孤立していると見られる写本群について、その本文の特徴的な因子（異文）を比較対照発見し、分析、整理することによって、しだいにその祖先の本文に近づきうると思われる。これにより原作者の原本への遡源も可能であると論理的

には考えられる。

一例として書写の系譜を考えて見よう。ここに甲乙丙丁戊己庚辛壬癸という１０本の写本があったとする。この１０本を校合した結果ＡＢＣＤＥＦａｂｃｄｅｆｇｈｉｊの１６個所の異文群が発見できたとする。この異文群が次図のようであったとすれば、甲乙丙の３本にはＡＢＣという共通本文があり、これが他の７本に比してより類似性が濃いと判断され甲１という共通祖本を持つと考えられる。同様にして丁戊己の３本は丁１を共通祖本に、庚辛壬の３本は庚１を共通祖本に持つと考えられ、３系統に整理される。この３系統本は、Ｃ・Ｄ・Ｅ・-の異文群を持つ。同様にして遡っていけば、甲２、甲３が想定されてくる。これに対して癸本は他の３系統の９本に対して、より多くの固有本文群を持っていることになるから、異本関係にあると言える。（以上本文批判に関しては、伊地知鉄男編著『日本古文書学提要』の“文献学”の項を参照した。）

図



以上のようなプロセスを経て、共通の祖本の本文に近づきうると考えられる。このような異文の調査に本文情報の活用が考えられる。

次に本文分析や作品分析の問題を考えて見よう。一つの作品の形象性を分析する際に、あるキーワードをさがしだし、キーワードを分析する手がかりを見つけるならば、きわめて有効な研究方法となる。この際Ａという作品から見つけ出されたキーワードをＢＣＤＥという作品におけるその語との用法上の違いを分析すれば、Ａ作品の性格の一部を明らかにすることは可能である。また、使用語彙を計量的に分析し作品を解釈、分析することも可能であろう。さらには文体的特質の分析から、作品分析

を行うことも可能であろう。こうした方法をとる際に本文情報、ことに語彙情報が活用されている。

3．索引誌作成の現状と問題点

次にテキストに対する情報がどのように活用されているかを見、かつかつそこにどのような問題点があるかを知るために索引誌作成の現状を分析する。

すでに、多くの語彙索引や術語、事項の索引が、文学研究や言語研究、歴史学の研究を目的として作成され公刊されている。これらは一部の例外を除いて手作業で作られている。この索引作成には膨大な労力が費やされている。これらの索引類を形式の面から整理してみると、

1）テキスト中に含まれるすべての語を対象として、その出現個所だけを示した総索引。→例1

2）テキスト中に含まれるすべての語を対象として、、見出し語をたて、その出現個所を示し、さらにその語を含む句または、文脈まで示した総索引。→例2

3）ある索引について複数テキストを対象として、またその中のすべての語を対象として、語の位置を示し、さらに文脈まで示した文脈付き総索引。→例3

4）ある作品中から特定の語（人名、地名等）のみをとり出し、その語のテキスト中における出現個所を示した部分索引。→例4

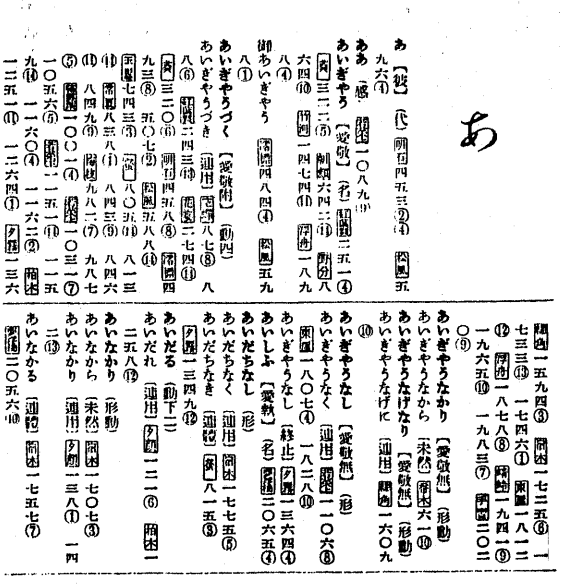
5）あるテキストについて、句を見出し語として、その出現個所を示した句索引。→例5

6）各種の資料（複数テキスト、複数作品）を対象として、術語をとり出し、その術語の含まれる文を掲出した事項索引。→例6

等になる。

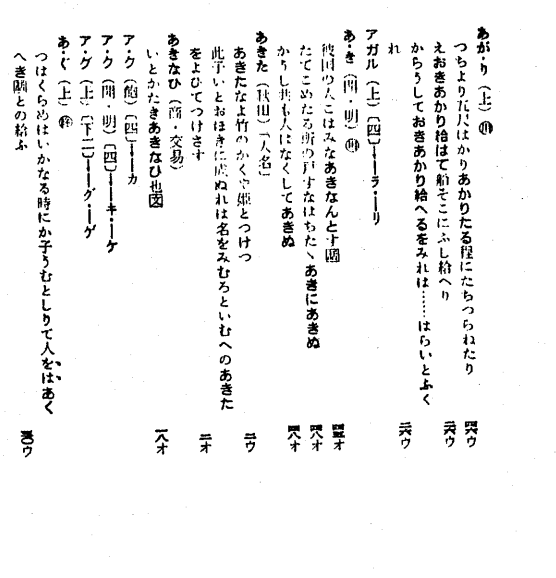
索引例1

（『源氏物語大成』より）



索引例2

（山田編『竹取物語総索引』より）



索引例3（上坂編『九本対照竹取翁物語語彙索引』より）

あ

あ（我・吾）（代名）（圏点）わ

あかほとけ　396（＊打仏㊀あるほとけ㊁㊂あるかほとけ㊃）

あ・う（逢）㊀㊁㊂㊃あふ

逢ことも涙にうかふ　500（あふ㊀㊁㊂㊃㊄㊅㊆㊇逢㊈）

あ・か（熊）㊀

かへり給はんことをあかすくちおしく覚しけれと　377

あか・さ（明）

ひるのあかさにもすきてひかりわたり　451（ナ㊀）

もち月のあかさを十あはせたるはかりにて　452

あか・し（明）㊀

夜をあかし日をくらすおほかり　23

ひとりあかしくらし給　232（明し㊁）

夜をあかし給へきにあらねは　382（明し㊁㊂）

あかし・の・はま（明石の浜）（地名）

はりまのあかしのはま也けり　257（明石のはま㊀あかしの濱㊃）

あが・ら（上）㊀

おきもあからてやひふせり　495

あが・り（上）㊀

つちより五尺はかりあかりたる程に　453（上り㊀）　＊あかる㊀＊めかる㊀

あはれ（哀）

これをあはれともみてをる　102（哀㊀㊁㊂）

すこしあはれとおほえけり　326（哀㊀）

＊むめる子のやうにあはれと　337㊀あれ

君をあはれとおもひいてける図　489（哀㊀㊁）　＊ころも㊁

あはれ・が・ら（哀）㊀

いといたくあはれからせ給て　497（哀から㊀）

あはれ・が・り（哀）㊀

月をあはれかり給へとも　392（哀かり㊀㊃）

あはれ・に（哀）㊀

せけんとゝろほそくあはれに侍る　395（哀に㊀㊁㊂㊃）

＊あはれか㊀

あ・ひ（逢）㊀

ひとり〳〵にあひたてまつり給ね　45

＊あひ給はむとおほす　48㊀あは

……といへはあひたてまつる　99

このみこにあひつかふまつり給へ　105（逢㊀）

かくや姫あひ給はす　210（逢㊁）

あひ・がた・し（逢難）㊀

あひかたしと思といふ　47（＊あひかたなし㊁）

あひ・かたな・し

＊あひかたなしとなんおもふ　47㊁あひかたし

索引例4（大隅編『大平記人名索引』より）

第Ⅰ部　日本人名（あ）

あ

アヤメ　→ 菖蒲ノ前

安芸ノ熊谷　→ 熊谷

安芸ノ勢
Ⅰ　227(7)
Ⅱ　135(13)，142(2)，383(4)
Ⅲ　122(9)

安芸前司　→ 島津安芸前司

安居院ノ中納言法印澄俊　→ 澄俊

安居院良憲法印　→ 良憲

安野中将公廉　→ 公廉

安保肥前守忠実　→ 忠実

安保肥前守忠冥　→ 忠実

阿保肥前入道僧禅　→ 忠実

安陪貞任　→ 貞任

安陪宗任　→ 宗任

安部視氏　→ 親氏

安間　→ 阿間

安間六郎左衛門　→ 利勝

安房ノ勢
Ⅰ　320(8)，332(12)
Ⅲ　185(14)

足立

索引例5

（馬淵編『今昔物語集文節索引巻29』より）

ア

アイギャウヅキタル（愛敬付タル）　二八　一八五　14地

形チ清氣ニテ愛敬付タル事ニテ似ス

アイギャウヅキタルガ（愛敬付タルカ）

清気ナル女ノ形チ愛敬付タルカ　三一　一三八　9地

アイギャウヅキテ（愛敬付テ）

年廿余許ノ女ノ下衆ナレトモ愛敬付テ糸清氣也　三三　一七六　8地

アイギャウヅ・ク（愛敬付ク）ーーキタル　ーキタルガ　ーキテ

アイシケル（愛シケル）

様ク愛シケル男子ノ　二七　一八三　8地

アイ・ス（愛ス）ーシケル　ースル

アイスル（愛スル）

極ハ子ヲ愛スル事世ヒモキ事ナレトモ　一一　一五八　12地

アイヨク（愛欲）ーーヲ

アイヨクヲ（愛欲ヲ）

愛欲ヲ発シテ婬タリ　三九　二〇六　9心

アカガネ（銅）ーーノ

アカガネノ（銅ノ）

銅ノ針其所ニ多クリ　一七　一六九　3地

索引例6（『古事類苑』文学の部より）

宇不足歌

〔宗長手記〕宗碩法師月村歌よみて文にあり中の七文字ふたもじたらず、

都には三十もじ餘一もじの二文字たらぬ歌も有けり、と中のぼせ侍り、

無同字歌

〔古今和歌集十八〕おなじもじなき歌　ものゝべのよしな

世のうきめ見えぬやまぢへいらむにはおもふひとこそほだしなりけれ

〔月明集下〕おなじもじかさならぬ

おのづから志りてあふげりすへくにはちとせたえせぬきみまさむとも

無濁音歌

〔玉勝間二〕濁る音なき歌

古今集に同じもじなき歌といふあり、そのたぐひにていはゞ、にごることなき歌もまゝあるもの也、音羽山おとにきゝつゝ相坂の關のこなたに年をふるかな、こむ世にもはやなりなゝむぬのまへにつれなき人を昔と思はむなど也、なほ有べし、

序歌

〔寄居歌談二〕萬葉集に、ものゝふのさつやたばさみ立むかひ射るまとかたは見るにさやけし、とあるは、則方はみるにさやけしのみに、歌のこゝろありて、上の十九字は、まとゝいふことをいはんがために、つくりかまへたる詞なり、これを序歌といふなり、これのみならず、萬葉古今のうちにいとおほし、江戸の仲田順忠がこひのうたに、

すみだ川洲にゐる鳥の志ほみてばせんかたもなくたつうきな哉

これらを分析してみると、次の点が指摘できるであろう。

1）近年公刊されている索引の多くは、比較的小さな作品ないしは中程度の作品に限られていることである。『源氏物語』の索引（例1）は、索引部だけで合計1，500ページ（3分冊）に及ぶものであるが、これが刊行されたのは昭和28年である。また『古事類苑』（例6）は、洋装本で四六判50冊。明治12年に建議され大正7年に完成したという。30年に及ぶ大事業であったが、これはほぼ60年前のことであった。

2）作品研究の進展は、一作品で一つのテキストのみならず複数テキストの索引を要求する。『竹取物語』研究のように、研究が進展し、より精緻な研究となっているところでは、例2、例3のように複数の索引が作られている。しかも例3のように九本のテキストを対照した索引までも作られている。

3）複数作られた索引を見ると、見出し語（索引語）のとり方が微妙に異っているのである。これは研究者（索引作成者）にとって、語の認定に差違が生ずることと、語の単位がその研究目的、方法、さらに対象によって異っていることを意味する。したがって一つの固定の索引では、すべての研究者の要求を満たせないことを意味している。

4．テキスト・データベースの必要性

以上のように、テキスト情報の活用方法、テキスト情報の活用上の一側面としての索引誌の現状を見てみると、単に語のデータベースを作るのではなく、テキストそのもののデータベース化が望まれる。このデータベースは、単に現在公刊されているような索引誌を作るベースとするだけではなく、語や句に対する要求、一文中の2～3語の組み合わせの検索、センテンスを検索する要求等、多様な要求に応えうること。また文学研究の多様な側面（本文批判や作品分析等）に応えうること。等をシステム的にサポートすることが必要であろう。

このためには、データベースとしてのごく一般的な条件の他に、語や文の認識をさせておくこと。本文以外の情報（例えば、品詞情報など）をも保持しうること。がテキスト・データベースの必要条件であると言えよう。これを可能とする蓄積構造を持つことが不可欠であると考える。

本研究は、昭和55～56年度文部省科学研究費試験研究（2）〔課題番号00581009〕「国文学情報検索システム及び索引誌の作成に関する研究」（研究代表者　市古貞次）の一部であることを付記する。