"Yoorei database" for dictionary compilation

『日本語科学』5（1999年4月）109-128

〔調査報告〕

国語辞典編集のための用例データベース

木村　睦子
加藤　安彦
田中　牧郎
（国立国語研究所）

キーワード

用例データベース，辞書，コンコーダンス，コーパス，KWIC

要　旨

既存の辞書に依存しないオリジナルな辞書を作るためには，まず生の用例をたくさん集める必要がある。国立国語研究所国語辞典編集室では，昭和54年に準備室が発足して以来，用例採集のための目録作り，作業手順の検討などを行うとともに，国定読本を資料として試験的に用例採集作業を開始した。昭和63年に正式の室になってからは，雑誌を対象に本格的な用例採集を開始し，続いて文学作品・国定算数教科書などに手をつけた。当面対象とする年代は，1901～1950年であるが，いずれは範囲を広げる予定である。

国定読本についてはすでにコンコーダンスが完成し，『国定読本用語総覧』1～12及びCD-ROMを刊行した（三省堂）。国定算数教科書についてもKWICと語彙表ができており，インターネット上での公開を考えている。現在最も力を注いでいるのが総合雑誌『太陽』であるが，これも3，4年のうちに，電子媒体で出版することになるだろう。文学作品については，多数の作家にわたるよう，短い作品をえらんで作業を進めている。

1．日本語用例辞典の構想

（1）歴史的国語大辞典であること

個々の見出し語について，意味用法・語形等の変遷がたどれるような辞書を作りたい。

（2）用例中心の辞書であること

歴史的変化をたどるとなると，内省というものは役にたたない。したがって，現代から上代に至るまでの用例を多数採集し，整理した上で収録する必要がある。ここに「日本大語誌」構想が生れた。当時はまだパソコンが普及しない時代であったから，書籍を念頭において，各時代ごとに○○巻，計164巻の用例集を作るという構想を立てた。現在は，媒体が何であるかはさておき，単なる用例集ではなく，用例集と辞書の一体化したものとして，用例辞典を考えている。意味記述は用例からの帰納を原則とする。

2．用例データベース作成の方法

2.1．全体構想

図「用例データベース作成の流れ」に示すように，用例採集の方法として，全数式（総索引方式）とスカウト式（選択採集方式）の二つを考え，対象資料の性質によって採集方法を使い分けるという方針をとってきた。すなわち，全数式においては，すべての用例を網羅して，見出しの頻度はもちろん，意味用法区分ごとに頻度をはかることができるようにする。スカウト式とは，人間が大量の資料に目を通し，採録したい語を抽出する方式をいい，異なり語数をふやすことと，めずらしい用法を拾うことを目的とする。

我々が手本として見ている外国語の大辞典にＯＥＤ（Oxford English Dictionary）とフランス語宝典（Trésor de la Langue Française 略称ＴＬＦ）とがある。ＯＥＤの用例採集法はすべてスカウト式であり，多数のボランティアの助けを借りた人海戦術によるものであった。採集した用例数は600万，辞書に記載した用例数が241万といわれる。他方ＴＬＦは，コンピュータによる全数調査方式をとり，１億の用例をもつという。選択採集による600万と全数式による１億とどちらが多いかの判定はむずかしい。スカウト式で600万語を採集するには，少なくとも１億語の文章に目を通す必要があるはずであり，ことによると大幅に上回っていたとも考えられる。この二つの方式の違いは，明らかに時代の差によるものであり，ＯＥＤ初版の時代には，人手によるしかなかったのである。

我々の方法は混合方式もしくは折衷方式である。初めは混合方式，すなわち，調査対象となる資料によって，全数式とスカウト式とを使い分けるという方法をとった。ＴＬＦより遅れてスタートしたにもかかわらず，漢字入力の問題があるために，フランス語の場合ほどコンピュータの威力が発揮できないと考えたのが，その理由である。とはいえ，わずか10年前と比べても，情報機器の発達はめざましく，この面で見直しをせまられている。最初の資料となった国定読本の文脈付き総索引作りは，昭和55年に手作業で開始し，途中から汎用計算機に切り換え，それからさらにパソコンに切り換えて，手作業でやった部分を作り直すという煩雑な手順をふむこととなった。また，最近は，全数式とスカウト式を折衷した代表例抽出索引方式を編み出し，採集方式を一貫して統合するコーパスの構想を打ち出している。

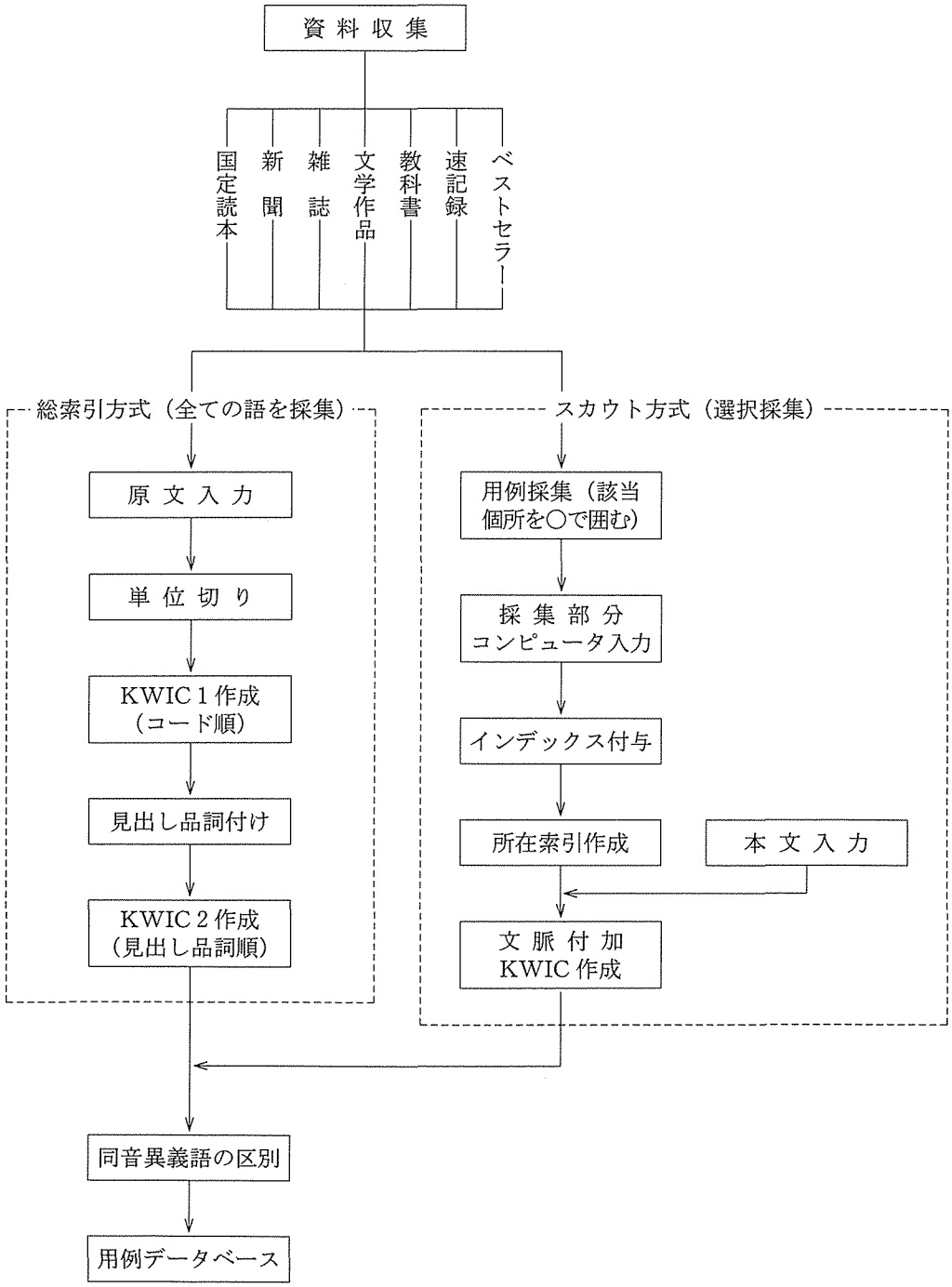
2.2．調査対象資料

用例採集の対象は，現代から始めて上代にさかのぼるという方針で，当面の目標を1901（明治34）年から1950（昭和25）年までとした。この時期をえらんだ理由は，その時期に日本語の標準語が成立したと思われるからである。その期間における雑誌，文学作品などの目録として，以下の三つができている。

用例採集のための主要文学作品目録	昭和58年	国語辞典編集準備資料２
用例採集のための主要雑誌目録	昭和58年	国語辞典編集準備資料３
用例採集のためのベストセラー目録	昭和59年	国語辞典編集準備資料４

これらの目録は，単に雑誌名や作品名をあげただけでなく，それらに語彙資料としての評点を

用例データベース作成の流れ



与えている。初めに網羅的な作品・雑誌等のリストを作り，それに評点を与えるために10人の評定委員を委嘱し，第一次リストから調査対象とすべきものを選定してもらった。ある作品について10人の委員のうち何人が票を入れたかによって，その作品の評点がきまる。

ただし，国語辞典編集室の規模からして多くの作業をこなすことは不可能であり，実際に手を付けたものは，以下の４種である。

1 国定読本（全数式）

2 国定算数教科書（全数式）

3 総合雑誌『太陽』（スカウト式）

4 文学作品（代表例抽出索引方式）

３．国定読本

3.1. 資料の性格

国定読本というのは，明治37年４月から昭和24年３月までの間に使用された文部省著作の小学校用国語教科書６種のことである。その６種を使用時期に従って示すと下記の通りである。

第１期　明治37年より使用『尋常小学読本』（通称「イエスシ読本」）８巻

第２期　明治43年より使用『尋常小学読本』（通称「ハタタコ読本」）12巻

第３期　大正７年より使用『尋常小学国語読本』（通称「ハナハト読本」）12巻

第４期　昭和８年より使用『小学国語読本』（通称「サクラ読本」）12巻

第５期　昭和16年より使用『ヨミカタ』『よみかた』『初等科国語』（通称「アサヒ読本」）12巻

第６期　昭和22年より使用『こくご』『国語』（通称「みんないいこ読本」）15巻

底本はおおむね初年度使用本である。

3.2. 作業経過

上記の本文すべてを単位切りし，各単位ごとに見出し・品詞・層別情報・文脈等をつけ，見出しの五十音順，品詞番号順などによって配列したのが『国定読本用語総覧』1～11（三省堂）である。文脈の範囲は人手によってきめたが，５期から一部（助詞・助動詞など）をＫＷＩＣ化して手間を省いた。これらの作業は１期から順におこなわれ，逐次書物として刊行された。そこで全体を見渡せるように，１～６期のすべての見出しに，期ごとの頻度と合計を付与し，語彙表の形にしたのが『国定読本用語総覧』12（総集編）である。さらにすこし遅れて，これらの用例をすべてＫＷＩＣ形式にして，CD-ROM で刊行した。

見出し総数３万２千，延べ用例数は60万である。

3.3. 用例データベースの構成

（１）書物と CD-ROM の違い

『国定読本用語総覧』は，書物と CD-ROM とで　①文脈の範囲と　②空見出しの有無　の２点において異なっている。すなわち書物においては，主として量的な制約から，すべての用例を均

等に扱うことをせず，人間が文脈の範囲を一つ一つ判断してきめたが，CD-ROMではすべての用例について，前後100字ずつ（キーは後文脈に含まれる）付与し，その範囲内で利用者が長さを指定できるようにした。また単位語が長単位であるため，書物においては検索の便を考え，後要素と称する空見出しを立てたが，CDにおいては部分一致検索ができるので，空見出しは入れなかった。それ以外の点はほぼ共通である。

（2）記載内容

データベースは見出し項目と用例項目とからなる。各項目のフォーマットは次の通りである。

〈見出し項目〉：〈見出し番号〉，〈見出し〉，〈漢字注記〉，〈品詞番号〉，〈品詞略号〉，〈同音語記号〉，
〈1期度数〉，……，〈6期度数〉

〈用例項目〉：〈見出し番号〉，〈出典番号〉，〈層別情報〉，〈前文脈〉，〈キー〉，〈後文脈〉

以下，必要な部分についてだけ注釈する。

① 見出し番号

見出し項目と用例項目とを結びつける唯一のものであり，他に共通部分はない。1～6期の全見出しを五十音順に配列し，5桁の一連番号を付け，末尾に0を加えて6桁にしたものである。語彙表の見出し番号と一致する。

② 漢字注記

原テキスト中での表記とは関係なく，語の識別のために付与したものであり，大体は『学研国語大辞典』によっている。語によっては漢字表記をもたないものもあるので，その場合には空欄になる。

課名・題名のように非常に長くなる可能性のあるもの，外来語を含む見出しには漢字注記を付けないのがきまりである。

③ 品詞番号・品詞略号

品詞番号と品詞略号とは一対一に対応するものなので，原理的にはいずれか一方でよいわけであるが，分かりやすさと機械処理の都合とから，二通り設けた。内容は以下の通りである。動詞を活用型によって分けたり，助詞を機能によって分類したり，通常の品詞より細かい区分になっている。

品詞番号	品詞略号	備考
01	課名	読本中の課の名称。番号も含む。
02	話手	せりふの上に記される話し手名
03	人名	歴史上の人物および架空の人物の名
04	地名	行政区画の他，山・川・海・砂漠等の名称
05	題名	課よりも下の区分の表題並びに一般の書名
08	名	上記以外の固有名詞を含む名詞一般
09	代名	代名詞

１０	形状	形状詞	７０	四	四段活用動詞
２０	副	副詞	７１	五	五段活用動詞
３０	連体	連体詞	７２	上二	上二段活用動詞
４０	接	接続詞	７３	上一	上一段活用動詞
５０	感	感動詞	７４	下二	下二段活用動詞
６１	格助	格助詞	７５	下一	下一段活用動詞
６２	副助	副助詞	７６	カ変	カ変動詞
６３	係助	係助詞	７７	サ変	サ変動詞
６４	接助	接続助詞	７８	ナ変	ナ変動詞
６５	並助	並立助詞	７９	ラ変	ラ変動詞
６６	準助	準体助詞	８０	形	形容詞
６７	終助	終助詞	９０	助動	助動詞
６８	間助	間投助詞	９９	（無品詞）	漢字の音訓やしりとりの文字など，意味のない文字列に与える品詞番号。略号はない。
６９	四五	同一語で四段活用（文語）と五段活用（口語）両方の用例があるもの。			

④　層別情報

本データベースにおける層別とは，資料の内容による分類ではなくて，文体などを指示するものである。種類は以下の３種である。

1　口語文　文語文　候文
2　散文　　韻文　　手紙文
3　地の文　会話文

上の各行ごとに一つずつを選ぶが，左端のもの，すなわち口語・散文・地の文は表示しない。同じ行の中から二つとることはしないので，候文の場合は「候」とだけ記して「文」とはしない。手紙の中に会話が引用されていれば「手会」と表示するが，手紙の中の和歌や俳句は単に「韻」または「文韻」と表示する。そして *1* ～ *3* の順に漢字１（または０）字ずつを層別情報欄に並べる。したがって最大３個である。

〔例〕013060,26206001038,文,首府の人口も年々著しく増加する,▼勢▽,なれば、其の巴里と同數に至るも亦甚だ遠からざるべし。

⑤　外字等

外字については，JIS 内の漢字で置き換えられるものは置き換えたが，それのできないものがいくつかあり，本文中にない記号で置き換えた。おどり字のうち JIS に含まれないもの，すなわち「くの字点」「二の字点」等は次のように表わす。

くの字点　　＠Ｋ１　　例：西へ＠Ｋ１（西へ西へ）
同上にごり　　＠Ｋ２　　例：それ＠Ｋ２（それぞれ）

二の字点　　@K3　　例：益@K3（益々）

⑥ ルビ

ルビを示すには［　］を用いる。たとえ熟字訓であっても一まとめにせず，必ず1字ごとに表示する。

例：田［い］舎［なか］

1字ごとに表示するのは，印刷の際にルビの形で出力するために必要なことである。

変体仮名は通常のひらがなに変え，前後を＊で挟んだ。

⑦ 検索ツール

検索ツールは用例検索のためにのみ存在する。語彙表や本文は汎用のエディタで扱える大きさなので，ユーザのディスクに移して自由に加工することができる。

検索ツールはプログラムと索引と語彙表とからなる。検索条件が与えられたら，まず語彙表を検索して条件に合う見出しを選んで見出し番号を抽出し，見出し番号によって用例データベースにアクセスする。その際に索引が必要なのは，用例データベースが約250MB と大きいためである。同一見出し語中での用例の選択は本プログラムでは行わないので，ユーザのファイルに出力してのち，他の手段を用いる必要がある。

検索プログラムは Borland社の Delphi Developer を下敷にしている。このソフトでは，日本語について，清音と濁音，拗音と直音の区別をしていない。区別をするのとしないのとにはそれぞれ一長一短がある。たとえば清濁の区別をしないことによって，「さかい」で検索しても「くにざかい」のように連濁したものも引き出せるし，旧かなづかいによる本文データを検索するには，「しょう」と「しよう」の区別をしない方が検索もれが少い。その代り，余分なものも多く出る。当面の検索対象は新かなづかいによる語彙表であり，原状のままで試験的に走らせたところ，ノイズがかなり目立ったので，区別をするように改めた。

4．国定算数教科書

4.1．資料の性格

国定算数教科書も読本と同じく1期から6期まであるが，1期は教師用のみで児童用がないので，2期以降を調査対象とした。底本として『日本教科書大系』（講談社）を用いた。この教科書大系は教科によって扱いが異なり，部分的に活字化しているところがあるが，算数はすべて写真版である。

算数であるから，当然練習問題がたくさんあるが，数式だけのものを除き，文章の体をなしている部分を調査対象とした。したがって読本より量が少なく，3分の1程度である。

第2期	『尋常小学算術書』	3～6学年	4巻	明治43．3
第3期	『尋常小学算術書』	3～6学年	4巻	大正 8．3
第4期	『尋常小学算術書』	1～6学年	12巻	昭和 9.12
第5期	『カズノホン』『初等科算数』	1～6学年	12巻	16．3
第6期	『さんすう』『算数』	1～6学年	9巻	22．3

4.2. 作業経過

作業手順は読本とほぼ同じであるが，単位の切り方や見出しの立て方，数詞の扱い方など，いくらか読本と異なるところがある。たとえば，読本では文語と口語で活用型の異なる動詞は別見出しになっているが，ここでは口語形に統一した。また数字（漢数字・アラビヤ数字とも）の列は，慣用表現を除き，ほとんどすべて〇に置き換えた（例：二割五分一〇わり〇ぶ，80糎一〇センチメートル）。これによって，助数詞はすべて語彙表の先頭にくることになる。

２～６期全体のＫＷＩＣと語彙表が一応完成している。ＫＷＩＣの文脈は短いが，出典番号が文字位置まで指示するようになっているので，任意の長さに作り替えることができる。今のところ出版の予定はないが，インターネット上で公開することを考えている。

4.3. データ量

1 原文　　1.1MB（約56万字）

2 単位切り本文　　1.4MB

3 語数（延べ・異なり）

	２期	３期	４期	５期	６期	全体
延べ語数	18,174	21,193	56,353	48,398	62,259	206,377
異なり語数	1,686	1,578	3,481	2,885	3,120	6,604

５．スカウト式と『太陽』コーパス

5.1. スカウト式と『太陽』

スカウト式による用例採集は，当初，国語辞典編集準備調査員であった，見坊豪紀氏が示した採集方法を範に計画を立てた。見坊氏の方法は，『スカウト式用例採集の手引き』（国語辞典編集準備資料８，昭和57年，以下『採集の手引き』とする），『スカウト方式による用例採集の実験的試行―「坊っちゃん」の場合―』（国語辞典編集準備資料９，昭和63年，以下『実験的試行』とする）に詳しく述べてある。本格的に採集を進めるようになってからも，この『採集の手引き』『実験的試行』を参照しながら，作業を行った。

『実験的試行』によれば，スカウト式とは，「あらかじめ設けた規準にかなう目標語だけを意図的に採集するやり方」であって，そのねらいは，

（１）全数調査にかける時間を節約し，

（２）それまでの調査で得られなかった新しい語形・用法を補充し，

（３）できれば，異なる語，異なる用法だけをすくい上げて，採集の能率化をはかる

ことにある（２頁）。つまり，全数調査では得られにくい語形や用法を効率よく採集することを目指したわけである。採集の着眼点について，『採集の手引き』では，Ａ：作品に即した採集，Ｂ：ことば・言語行動に関係のあるもの，Ｃ：単純語より大きいことば・小さいことば，Ｄ：語形上の着眼点，Ｅ：さまざまの用語，Ｆ：意味・用法，Ｇ：用字・表記，Ｈ：誤り，の八種にわたっ

て細かく列挙している。ただし，着眼の規準は客観的に示されておらず，採集者が経験的に習得していくべきものである。

前述の通り，国語辞典編集室の用例採集事業は，全数式とスカウト式の二本立てで進めてきた。資料の性質によって採集方式を変えることで，効率的に多くの用例を集めることをねらったのである。スカウト式の対象に予定したものには，新聞，雑誌，速記録，ベストセラー，代表的作品以外の文学，読本以外の教科書，などがある。全数式の対象とした，国定読本，代表的文学作品を補うべきものとして，スカウト式による資料を位置付けた。1901～50年で3,000万の用例を採集する目標を立て，そのうち2,000万例をスカウト式によって，さらにそのうち1,000万例は雑誌から採集するという計画であった（『実験的試行』１頁）。つまり，用例の数から見れば，スカウト式による雑誌からの採集が３分の１という大きな部分を占めるわけである。採集対象にする雑誌は，評定委員10名の推薦に基づいて選定した。10名中４名以上が推薦した雑誌120誌がリストアップされたが（『用例採集のための主要雑誌目録』，昭和58年），そのうち10名全員が推薦したのが，『太陽』『改造』『文芸春秋』『婦人公論』『子供の科学』『帝国文学』『アララギ』『ホトトギス』の８点であった。この８点のなかから，広範囲の内容を覆える総合雑誌であり，国立国語研究所が全巻を所蔵しており便利であるなどの理由から，『太陽』が最初の採集資料として選ばれたのである。はじめに，当面対象とする時期の始まりとしての1901年から８年ごとに，1909年，1917年，1925年の各12冊計48冊を定め，これに，『太陽』としての完結性をもたせるため，創刊年1895年の12冊と，終刊年1928年（２号で終刊）の２冊を加えて，合計62冊を対象に選定した。

5.2. 作業の手順

『太陽』に対するスカウト式用例採集は，具体的には次のような７つの段階を順次進めてきた。ただし，この７つの段階は当初の計画にはなく，作業を進めながら逐次予算を獲得し，手順を構築していったものである。特に④以後は，ここ数年になって手順に加えたものである。

① スカウト

スカウトを担当したのは，国語辞典編集室の非常勤研究員・通信研究員，計26名である。ほとんどは，国語関係の研究者または大学院生である。『採集の手引き』を参照のうえ，原文のコピーに対して，採集すべき語に赤丸を付ける手法で進めた。

② 採集語のパソコン入力

当初は，カードによる用例の蓄積と利用を前提にしていたが，用例の管理と検索を容易にすべく，パソコンを導入した。スカウトによって資料に赤丸が付けられた部分を，所在コードとともにパソコンに入力した。赤丸が付けられた部分を含んで，ある程度の長さの文字列を入力したが，その単位は，特に統一を図らなかった。なお，この段階から後は，『採集の手引き』『実験的試行』には記していない。

③ 採集語への読み付け

②の採集語に対して，読みを付与した。読みの単位等，読み付けにあたっての規準は，②の方式とともに，『スカウト式用例採集処理の手引き』（国語辞典編集準備資料11，平成７年）にまとめた。

④　採集語への文脈付与

採集語を用例として使えるものにするには，やはり，ある程度長い文脈をもっていることが望ましい。②の文字列では，文脈としては短か過ぎるので，別に入力が必要になる。１行に１例程度の高密度でスカウトされた『太陽』の場合，個々の採集語に文脈を入力していくよりは，本文の全文をまとめて入力し，プログラムによって，機械的に文脈を付与する方が効率的である。科学研究費：新プロ「日本語」(研究代表者：水谷修，平成６～10年度）の研究班４「情報発信のための言語資源の整備」の一環として，予算的措置に見通しがついたこともあって，本文入力を行うことで文脈を付与することにした。

⑤ 本文作成

本文の全文入力にあたっては，原資料の本文批判が必要になった。外字や異体字，仮名遣い，語法や用語・用字のゆれなどに関して，当時の言語状況を踏まえた処理規準をマニュアルに定め，これに則って本文を作成していった。こうして得られた本文は，それ自体を本文データとして利用することも可能になった。

⑥　採集語のキー位置指定

採集語の読みをもとにして，計算機を用いて本文の当該箇所の文脈を引くには，採集語と本文とをマッチさせるキー位置の指定が必要になる。このキーに相当するものとして，②で入力した文字列を修正して利用することにした。採集語の語頭から数文字を，キーとしてインデックスファイルに格納しておくことで，キーと所在コードとによって，本文から当該箇所の文脈を引き出すことができるようになった。

⑦　採集語への情報付与

採集語に対しては，読みだけでなく，漢字注記，品詞，語種，備考を付与する。これらの情報は，語の同定のために必要であるばかりでなく，多様な検索を行うためにも有用である。情報付与のための規準を定め，これにしたがって作業を進めた。

以上７つの段階のうち，①②③は，平成８年度までに，62冊すべてについて作業が終了している。全文入力を進めるように方針を変更した平成７年度から，④⑤⑥⑦に着手し，現在も継続中である。1901年12冊分については，⑦まで一通りの作業が終了しており，残る50冊分についても，あと３年程度で終了の見込みである。

5.3. 採集語の概要

スカウト式を導入するにあたって，この方式によってどのような語が採集されるかについて実験が試みられたことがある（高梨信博「スカウト方式による用例採集法の実験について」『研究報告集』５，国立国語研究所報告79，昭和59年)。そこでは，スカウト式のさまざまな問題点を指摘した上で，「あらためて，実験試行がつみかさねられるべきである」(124頁）としている。今回，一通りの作業が終了している1901年12冊分のデータによって，スカウト式によって採集された語がどのような性質のものであるかについて，品詞と語種の観点から概観してみたい。また，採集語が異なり語のどの程度を網羅しているのかについても，探ってみたいと思う。

スカウト式による『太陽』1901年分の採集語を，その読みによって数えると，異なり約61,000語，延べ約155,000語となる。ただし，この読みは，5.2.に述べた③の段階で付けられたもので，前方一致検索のために後要素が切り出されたり，参照読み（熟字訓に対する字音読み，音訛形に対する規範形）が付けられたりしており，一つの採集語に複数の読みが付けられていることも多い。こうした付加的な読みを除き，さらに，助詞・助動詞，連語を除いたものについて，品詞構成を整理すると次のようになる。

『太陽』1901 年分　スカウト式採集語の品詞構成

	異なり	%	延　べ	%
体	37,728	74.0	67,058	63.3
用	8,372	16.4	24,519	23.1
相	4,631	9.1	13,445	12.7
その他	231	0.5	943	0.9
計	50,962	100.0	105,965	100.0

『雑誌用語の変遷』（国立国語研究所報告89，昭和62年）で調査された『中央公論』のデータと比較するために，ここでの品詞枠は原則として，『分類語彙表』の４分類にしたがう。ただし，『分類語彙表』の枠組では，陳述副詞は「その他」に含まれるが，『太陽』の品詞情報では副詞に区別を与えていないため，陳述副詞も「相」に含まれている。次に，『中央公論』のデータのうち，1901年とほぼ同時期の1906年の品詞構成を掲げる。

『中央公論』1906 年分　抽出 10,000 語の品詞構成

	異なり	%	延　べ	%
体	2,885	64.4	5,102	51.0
用	932	20.8	2,913	29.1
相	579	12.9	1,606	16.1
その他	87	1.9	379	3.8
計	4,483	100.0	10,000	100.0

これは，標本抽出によって，延べ10,000語を取り出して調査したものである。なお，『中央公論』の調査では〈長い単位〉を採用しており，『太陽』の読みも基本的にこの単位を踏襲している。この二つの表の比較から明らかなことは，スカウト式による『太陽』は，標本抽出による『中央公論』よりも，体の比率が高く，用・相・その他の比率が低い，ということである。ほぼ同時期の同種の資料に，こうした差異が見られるのは，採集抽出の方法の異なりによるものと考えられる。スカウト式では，体の類が採集されやすく，用・相・その他の類は採集されにくい，という傾向が確かめられる。

次に，同じデータについて，語種構成を比較してみると，次のようになる。

『太陽』1901年分　スカウト式採集語の語種構成

	異なり	%	延　べ	%
和　語	8,642	17.0	25,981	24.5
漢　語	31,686	62.2	59,322	56.0
外来語	1,427	2.8	2,184	2.1
混種語	9,207	18.0	18,478	17.4
計	50,962	100.0	105,965	100.0

『中央公論』1906年分　抽出10,000語の語種構成

	異なり	%	延　べ	%
和　語	1,595	35.6	5,799	58.0
漢　語	2,189	48.8	3,260	32.6
外来語	72	1.6	87	0.9
混種語	627	14.0	854	8.5
計	4,483	100.0	10,000	100.0

ここから，スカウト式による『太陽』の方が，標本抽出の『中央公論』よりも，漢語・外来語・混種語の比率が高く，和語の比率が低い，という傾向がとらえられる。スカウト式では，漢語・外来語・混種語が採集されやすく和語は採集されにくいことが確かめられる。

品詞構成や語種構成に見られる，スカウト式による採集語の特徴は，おそらく，品詞や語種に反映するある性格に基づくものであろう。用・相・その他や和語には，高頻度で基本的な語が多いのに対して，体や漢語・外来語・混種語には低頻度で周辺的な語も多い。全数調査では得られにくい語を拾うことを着眼点の一つとして異なり語を増やしていこうとするスカウト式の趣旨に沿った結果であると考えられる。

このように，異なり語を増やしていこうという目的にかなった採集語が得られていると考えられるが，いったい，どの程度の異なり語を網羅しているのだろうか。このことを知るために，『太陽』の1901年第２号１冊を取り上げて，簡易全数式によって抽出された異なり語とスカウト式によって採集された語とを比較してみることで，スカウト式用例採集の網羅性をはかってみた。ここでいう簡易全数式による異なり語の抽出とは，次のような方式によるものである。まず，電子化した全文テキストに対して，機械的に漢字・かな等の字種を目印に区切りを入れ，手作業によって一定の単位に修正を加える。切り出された単位に見出しを与え，同一の見出しをもつものが複数ある場合は，ひとつだけを残して他は消去する。この簡易全数式によって得られた異なり語を，スカウト式による採集語と比較するわけだが，単位や規準をそろえるために若干の操作を加えた。

例えば，スカウト式で，後要素や参照読みなど，一つの採集語に複数の読みが付けられている場合は，いずれか一つの読みが，簡易全数式の異なり語の見出しと重なっていれば，その語は一致した例と見なす，などの操作である。

調査の結果は，次の通りである。

スカウト式の採集語と一致するもの　　7,502語　　45.5%

スカウト式の採集語と一致しないもの　　8,985語　　54.5%

この調査によれば，スカウト式によって採集された語は，異なり語全体の半数弱しか網羅していないことになる。しかし，これは62冊のうちの１冊についての値である。つまり，全体で１，２例しかないような希少例は別として，他の61冊のどこかに用例があれば，そこで拾われる可能性はある。とはいえ，異なり語全体の中で低頻度語の占める割合はかなり大きなもので，これらが網からもれる恐れは少なくない。その意味で，本文データができ全文検索ができるようになったことは，用例データベースを辞典編集に利用すべき立場にある者にとって喜ばしいことである。

5.4.『太陽』コーパスの作成

5.2.に述べたように，『太陽』については，スカウト式によって採集語を集積していく形態から，本文データとしても利用できる形態へと，その目指すところが変わってきた。本文全文から多様な検索方法によって用例を自在に引き出せるようなコーパスの作成を目指すのである。その実現のためには，信頼できる本文を提供することと同時に，検索性の高いインデックスを整備することが不可欠である。

『太陽』は，総合雑誌という性格から，非常に広範囲のジャンルを含んでおり，単一の資料でありながら，多層的な価値をもっている。1901年分の本文データにおけるジャンル別文字数は，下表の通りである。

1 論　説	417,939	9 歴史地理	427,327	17 家　庭	187,657
2 政　治	106,302	10 伝　記	82,184	18 投　書	20,824
3 経　済	272,909	11 随　筆	38,947	19 編　集	20,101
4 法　律	74,408	12 文　芸	146,813	20 談　話	115,906
5 教　育	60,551	13 社　会	60,064	21 小　説	290,175
6 宗　教	55,514	14 海　外	162,961	22 韻　文	16,760
7 農工業	301,777	15 世　論	98,415	23 漢　文	77,705
8 科　学	106,618	16 彙　報	176,100	24 その他	385
				合　計	3,318,342

このジャンル別分類の枠組みは，原文の欄別等を指標にしたもので，その対応は次の通りである。（　）の中が欄の名称，〔　〕の中は欄の名称ではないが内容から規定できるものである。

１論説（論説・太陽〔巻頭言〕），２政治（政治時評），３経済（経済時評・商業世界），４法律（法律

時評），５教育（教育時評），６宗教（宗教時評），７農工業（農業世界・工業世界），８科学（科学世界），９歴史地理（歴史地理），10伝記（人物月旦），11随筆（小説雑俎〔随筆〕），12文芸（文芸時評），13社会（社会時評），14海外（海外事情），15世論（与論一般），16彙報（海内彙報），17家庭（家庭談叢），18投書（寄書），19編集〔編集部記事〕，20談話（名家談叢），21小説（小説雑俎〔小説〕），22韻文〔詩・短歌・俳句〕，23漢文〔漢文〕，24その他〔皇室関連〕

本文データに対して，こうしたジャンル別情報を与えておけば，ジャンルを限った検索が可能になるばかりでなく，用例のジャンルによる分布などを一目で知ることもできる。例えば，このジャンル別情報なども，検索性を高めるインデックスの一つとして，有用性の高いものであろう。

検索性の高さを求める場合，文字列検索だけでは不十分で，テキストに対して何らかのインデックスを付与することが不可欠である。全文を単語や形態素に区切って読みや品詞等の情報を付与する，全数式によるインデックスの付与が望まれ，実際に，『太陽』の本文の一部に対してはこれを試行している。しかし，大規模な本文全体に対してこれを及ぼすことは，労力の観点から見て現実的ではない。こうした全数式のインデックス付与と，もっと簡便な他の方式によるインデックス付与との共存を図るべきであろう。スカウト式による採集語は，そのような方式によるインデックスの一つとしてとらえることができよう。採集語そのものを引き出すインデックスとしてだけでなく，採集語を手がかりに本文データ全体に対して改めて検索をかけることができるシステムとして，練り直すのである。たとえば，動詞「とる」に該当する例を多数採集したいとき，採集語から「取る／執る／採る／資る／摂る／摑る／捕る／娶る／とる」などの表記を得，これらの表記によって文字列検索をかけるのである。一定の方針に基づいて，人力によって選択された語は，その本文全体を何らかの面で代表していると考えられる。スカウト式によるインデックスが，他の方式のインデックスや文字列検索などと補い合って検索性を高めることができれば，『太陽』コーパスの価値は格段に高まるであろう。そして，用例採集方法としてのスカウト式にも，新たな価値が見出だされるであろう。

なお，『太陽』コーパスを公開するにあたっては，著作権の問題が残されている。次に掲げる表は，『太陽』1901年分全記事の著者の著作権について，没後50年を経過しているか否かでその状況を整理したものである。

署名記事	著者没後50年以上経過（1999年現在）		８１６本
	著者没後50年未満	2002年までに没後50年	９５本
		2003年以後に没後50年	７１本
	著者没年等の情報不明		１１８本
無署名記事　（匿名記事・編集部記事含む）			５２４本
全記事			１６２４本

無署名記事については，刊行後50年で著作権が消滅するから，雑誌終刊の1928年から50年以上経過していることで問題はないであろう。署名記事については著者の没後50年までは著作権が保

護されている。そこで，1901年分については網掛け部分（200本弱）がコーパスを公開するには問題となる。この状況のままでは全体の一割強が著作権をクリアできないことになり，1909年以後については，その割合が高くなるものと思われる。没後50年を経過していない場合は，著作権者に承諾を求めていきたい。署名されていてもその著者の没年等が不明なものについては，調査文献や調査機関を広げるなどして解明に努力したい。

『太陽』コーパスは，３～４年後に電子媒体によって公開し，一般の利用に供する予定である。

６．代表例抽出索引方式と文学作品

6.1．代表例抽出索引方式の位置付け

代表例抽出索引方式による用例採集は，国語辞典編集室におけるそれ以前の全数式とスカウト式という２種類の用例採集方式の利点を生かすことを特徴として平成６年度より始めた。

国語辞典編集準備室開設当初はカードを用いた手作業による用例採集を行う計画であったが，計算機のデータ処理能力の向上によってそうした用例採集を計算機を用いて行うことが容易になってきたため，代表例抽出索引方式では最初から採集作業の流れ全体に計算機（パソコン）が介在する方式として計画した。基本的には，全数式と同じく対象資料全体の語を扱う方式である。

全数式用例採集の特徴として，

・対象資料全体の語の種類
・対象資料中の全語数
・品詞毎の語数やその構成
・同一語の異表記のパターン

など，資料全体にわたる情報の把握が容易である点をあげることができる。

また，スカウト式用例採集の特徴としては，5.1.に述べたように，全数式による用例採集に比して時間的な節約が可能な点をあげることができる。

両者の特徴を裏返せば，全数式は対象資料全体の語を採集対象とするために短期間で大量のテキストデータを扱うことが難しく，スカウト式は短期間に全数式よりも幅広くより多くの資料にあたることや異なり語数を増やすことは可能であるが，出現した語全体を扱えないために対象資料を語彙という視点から特徴付けて把握することが難しいということである。そうした事情から，全体のデータ量が少ない国定読本に全数式が採用され，内容が多方面，多岐にわたる雑誌『太陽』にスカウト式が採用されたわけであるが，代表例抽出索引方式は，全数式と同様に対象資料全体の語を対象としながらもスカウト式と同様に短い期間で用例採集ができるよう，両者の中間的な文章形態の文学作品を対象とした。

6.2．代表例抽出索引方式の採集作業

では，具体的にこの方式による用例採集作業の流れについて述べる。

（１）対象資料

当初対象と考えた文学作品は，『用例採集のための主要文学作品目録』（昭和55年，以下『主要文

学作品目録』とする）にあげられている1869（明治２）年から1966（昭和41）年までの期間の139作品のうち，現在国語辞典編集室で用例採集の対象時期としている1901～1950年の50年間に発表された文学作品116点である。

その後，116点の作品のうち１万文節以上の作品は対象から外し，さまざまな作家の１万文節以下の作品から幅広く偏りのない用例採集を行うよう若干の方針変更を行った。これは長編の多い作家あるいは『主要文学作品目録』に長編のみあげられている作家の場合，どうしても採集結果にその作家の使用語彙への偏向が出てしまい，延べ語数の増加に比して異なり語数の増加が伴わない結果になるのではないかと考えたからである。また，１万文節以下の短編とするのは文学作品中の同一語，特に固有名詞（登場人物名，地名など）の重複をできるだけ押さえたいということと，作品の背景の違いが使用語彙における異なり語数の増加につながると考えたからである。異なり語数の増加にこだわるのは「代表例抽出索引方式」という方式のもつ特性に関係している。それは，この方式が最終的な形態として，個々の語の代表的な意味・用法に重きをおいた用例集を目指しており，既に収集した用例と比較して意味・用法が同一のものはできる限り用例集からは外すことを考えているからである。より多くの作家と作品から，語はもちろん，意味・用法の異なりを多く獲得することもこの方式の目指すところである。

これらの文学作品の選定は，以下の手順に従って行われた。

まず，現代の代表的な文学全集15種類に収録されている全作品をリストアップし，それをもとにして「主要文学全集収録作品目録」を作成（『主要文学作品目録』第Ⅱ部に掲げられている）。次に，「主要文学全集収録作品目録」に収められた全1506点の作品について，近代文学研究の専門家及び作家10名に100作品程度ずつ推薦を依頼した。

このとき，口語体，文語体の別は問わなかったが，

①著名な作家の代表的な作品，文芸的価値の高いもの。

②その時代によく読まれ，または現代まで読み継がれてきたもの。

③用語文章が標準的で，一般読者への影響があったと考えられるもの。

ということを作品の性格としてもつものと限定した。

また，年代や作家に偏りが出ないこと，初版本の入手しやすいもの，という点も考慮した。

この結果，1506点の作品のうち，572作品が選定された。さらにこの572作品の中で選定者４名以上からの推薦を受けた作品139作品を主要な文学作品とした。

年代別に４名以上の方が選定された作品数をみてみると，

1901年～1950年：116作品

1868年～1900年： 17作品

1951年以降： 6作品

となっている。

その後，作業を進めるうち，当面の作業対象文学作品について，

・より幅広く語の異なり用例を集める

・作家／作品による特定の語の出現数偏向を避ける

という観点から，116点の作品のうち，概算で１万文節以下のものを当面の対象とし，さらに『主要文学作品目録』作成の準備段階資料である「用例採集のための文学作品選定目録」から，やはり概算で１万文節以下の作品を優先して対象とするように方針に変更を加えた。

以下に現在の対象文学作品数を示す。

「用例採集のための主要文学作品目録」	46作品	１万文節以下45作品 １万分節以上 1作品
「主要文学全集収録作品目録」	205作品	１万文節以下※
	83作品	１万文節以下の作品のない作家の最も短い作品
合　　計	334作品	

※ 目録上は 206 作品あげられているが，実際には文節数概算の誤りが１点あり，それを除いた数。

次に対象文学作品の種類と作品数及び概算文節数を示す。

種　類	作品数	文 節 数
小　説	280	1,490,580
評　論	34	105,430
随　筆	8	16,690
戯　曲	8	46,290
童　話	4	10,780
合　計	334	1,669,770

※　文節数で作品の規模を示すのは，助詞・助動詞を除いた延べ語数に近い数がそれで得られると考えたからである。

これらの対象となる文学作品全体については，書誌調査を行い，できる限り初版を底本とすることを基本方針としている。初版本そのものの入手は目的ではないので，初版の所在や復刻版の有無を確認して，コピーあるいは復刻版を入手し，それらを電子化してデータ作成を行うこととしている。初版が見つからない，あるいは複写不可の場合，書誌調査結果で判明した初版から５年以内に出版された単行本を使用することにしている。また，単行本がない場合は所載の雑誌を底本データとして利用することにしているが，初出の情報についてまだ不明なものが30作品余り残っている。また，著作権についても『太陽』の場合と同じように調査を進めている。

（２）採集作業の流れ

作業方法としては国定読本に対して行われた全数式用例採集作業とほぼ同様の手法で採集作業を進めていく。以下は書誌調査によって初版本もしくはそれに準ずる底本とするに足るテキストを入手した後の流れである。

①　本文入力

本文入力は，当初室員による入力，OCR（光学読取装置）を利用しての入力など試みてみたが，１文字当たりの入力単価が下がったため，現在は外部業者への発注によって入力を行っている。

② 本文チェック

外部で入力されたテキストデータの校正を行う。このとき，漢字によってはJISコード上にないものがあり，それらはできるだけ国語辞典編集室内で取り決めた包摂規準によって代替漢字を用いるなどして処理を行う。それができない場合はゲタ記号（＝）によって置き換える。また，かなについても変体仮名のようなものなどについてはその情報を記録して通常のかなを用いて処理する。

③ 単位切り

テキストデータの本文校正が終了したものを，一定の単位（最終的な用例集における見出し語）によって切り出す。

④ ＫＷＩＣ作成

一定の単位に切られた本文データをプログラムにかけ，各単位毎にＫＷＩＣすなわち切り出された単位（語）の前後の文脈が見えるようなデータ一覧を作成する。この時点で単位に不揃いなものがあれば単位の切り直しや字句の修正を行う。

⑤ 情報付与

③，④の作業を終えたデータに対して，文脈を参照しながら各見出し語の読み，出現形，品詞，層別情報などの情報を付与していく。

現在，繰り返しテキストデータへのチェックが行われた50作品余りの本文データが作成されている。上記作業と並行して，『太陽』コーパスの全数式によるインデックスとの統一的な単位の規準，旧漢字の処理についての規準などの検討が進んでおり，データの構成そのものも『太陽』コーパスを含めた今後のデータ作りに齟齬を来さないようにしていこうとしている。

７．今後のデータ作成：国語辞典編集室コーパスの作成

これまで，日本語用例辞典の構想のもと，1901～1950年に限って，対象とする資料の種類や作業環境に応じて，全数式，スカウト式，代表例抽出索引方式といった，いくつかの方式で用例採集を進めてきた。一方，ある時期（例えば1901～1950年）における語彙，意味，文法といったことに言及し，それぞれの特徴を把握するような研究のためには十分な量のデータが必要である。多数のテキストを電子化して蓄積し，必要ならばそれらのテキストに含まれる語それぞれに統語情報や意味情報，あるいは表記などの情報を付与したデータの集合体，すなわちコーパスと呼ばれているものが，そのために最もふさわしいデータの形態である。これまで用例採集として進めてきた国定教科書（読本，算数），雑誌『太陽』などの用例データベース，文学作品や今後対象とする資料のデータ全体を，ひとつのコーパスとして統括することでそうした研究に資することが可能となる。もちろん，一口にコーパスといってもさまざまな種類があり，国語辞典編集室コーパスとして提供できるのは，対象資料の本文からなる生コーパスと，種々の情報を付与したタグ付きコーパスの形態のものの２種類である。

以下に今後の国語辞典編集室コーパス作成の具体的な方法とアイデアについて述べる。

・書かれた資料を計算機で読めるよう，テキストの電子化を行う（OCR または外部発注）
・電子化テキスト上の文字などをあらかじめ設けた規準にしたがって加工・処理する（JIS コード上にない文字などの処理）
・作成された本文データを形態素を基本とする短い単位に分割（第１次単位）
・分割されたそれぞれの単位をもとにＫＷＩＣを作成
・ＫＷＩＣデータを参照しながら，それぞれの単位に必要とされる見出し語を付与（第２次単位）
・見出し語に対して読み，漢字注記，品詞などの情報を付与
・さらに見出し語が前後の語と結合してより大きな単位となるものすべてをコーパスへの検索のキーワードとして付与（第３次単位）

今後作成される生コーパスについていえば，データには年代，出典，ページ，行などといった対象資料自体の情報や本文上の所在を示す管理情報が必要であり，また，テキスト上の JIS コード外の漢字や特殊な文字列の処理方法を検討していかねばならない。タグ付きコーパスについていえば，付与すべき情報，例えば，ここで仮に第１次～第３次とした単位各々についての規準や，複数の単位に対応できるような検索用キーワードに関しての規準を検討していく必要がある。また，データのさまざまな加工段階に応じた共通作業マニュアルの整備も必要で，これまでに蓄積されてきたデータをもとにそうした規準についての検討を続けているところである。

将来，こうしてコーパスが作成され，データが十分な量となった段階で，コーパスからのデータ抽出を行って「日本語用例大辞典」の編集を始めたいと考えている。

付　記

本稿は，国立国語研究所創立50周年記念研究発表会（平成10年12月14，15日）における口頭発表，「国語辞典編集のための用例データベース」の一部をもとに，まとめ直したものである。

（投稿受理日：1999年２月24日）

木村　睦子（きむら　むつこ）
国立国語研究所国語辞典編集室　115-8620　東京都北区西が丘3-9-14
加藤　安彦（かとう　やすひこ）
国立国語研究所国語辞典編集室
kateaux@kokken.go.jp
田中　牧郎（たなか　まきろう）
国立国語研究所国語辞典編集室
mtanaka@kokken.go.jp

Japanese Linguistics 5 (April, 1999) 109-128 [Article]

"Yoorei database" for dictionary compilation

KIMURA Mutsuko, KATO Yasuhiko, TANAKA Makiro
The National Language Research Institute

Keywords

yoorei database, dictionary, concordance, corpus, KWIC

Abstract

We are building a "yoorei database" (assembled concordances) for the purpose of compiling an original dictionary that is not dependent upon existing dictionaries. We assume that standard Japanese was established between 1901 and 1950. We set this period as our starting point to build our "yoorei database". In order to cover a wide range of our database objects, we take into consideration such materials as school textbooks, newspapers, magazines, literary works, best selling books, and the Diet Record. Three sets of materials that have already been worked on follow.

1. Japanese school textbooks written by the Ministry of Education (Completed)
   "Kokutei Tokuhon Yoogo Sooran" (A concordance of Kokutei Tokuhon) vol. 1-12 and CD-ROM version. Total number of words: 600,000 Total number of different words: 32,000 These numbers were counted by the "zensuu-shiki" method (to make a concordance of whole words from a data set).
2. Popular magazine "Taiyoo" 62 volumes (In progress. Planned to be done within 3-4 years.)
   Total number of words: 550,000 by the "scout-shiki" method (to make a concordance of selected words by "scout standard" from a data set). We are inputting the text data of the entire 62 volumes in order to provide contexts for the materials. We estimate that this text data set will consist of 8,000,000 characters.
3. Literary works: 334 stories (In progress)
   We chose 334 various short stories (including novels, essays, critical essays, drama and juvenile stories) from the point of view of creating a well-balanced concordance. We estimate the total number of words at 1,700,000, excluding postpositional particles and auxiliary verbs.
4. Arithmetic school textbooks written by the Ministry of Education (Completed)
   KWIC (Key Word In Context) data and the lexicon from this material have already been done. (Unpublished) Total number of words: 200,000 Total number of different words: 6,600

We have tentatively tried different methods for each data set. At this point, however, we are planning to make a manual (for building the "yoorei database") more complete in order to build the database using the most effective methods.