用例データベース作成のための単位と用例データベースの使い方

用例データベース作成のための単位と用例データベースの使い方

国語辞典編集室　　加藤 安彦（kateaux@kokken.go.jp）

要旨：

ここでは、現在用例採集作業が進行中の 1901 年～1950 年までの 50 年間における雑誌『太陽』及び文学作品のデータに対して考えている単位の付与のしかたについて述べる。

ここでの単位とは、

○ 用例を収集してデータベース化した時点で検索などにおいて表面的に現れる語

○ 用例データベースから欲しい用例を検索・抽出するきっかけとなるキーワード

などを考える上で必要とされるもので、「個々の語の機能や意味などの情報を担う用例データベースでの最小のかたまり」といった意味である。

今後の用例採集においてどのような長さ・種類の単位を考えていくべきか、その情報をどのようにデータベース上に格納していくべきかについて述べる。なお、この用例データベースは将来的にデータ量を拡大していき、その情報をもとにコーパス化することを考えているもので、その意味でも単位についてゆらぎの少ない規準作りが必要となってくる。

また、最後にこうして作成された用例集をどのように利用していくか、その簡単な例を挙げる。例として使用するのは、既に用例採集作業を終えた第一期から第六期までの国定読本データで、これには冊子体の他、CD-ROM 版でも公刊されているので、その CD-ROM 版によって説明を行う。

キーワード：用例データベース、コーパス、用例採集、国語辞典、単位

１．はじめに

日本語の書きことばでは、英語、フランス語などのように語と語の間にスペースを入れることをしない。その結果として、原稿や記事などの大きさを言うときに、欧米でよく使われるような全体で何語という数え方はせず、全体で何字といった計算をする。何字、ということを知ることは、その原稿や記事全体の大きさを考えるのにはよい手がかりである。何語、という計算方法だと、" staphylococcus（ぶどう球菌）"などといった長い語が頻出するような文章では全体量がどれほどのものになるか想像しにくい。しかし、文字数だけでは、ある作家の小説一作品がどのくらいの語から成り立っているのかを知りたいときには不便で、結局書かれたことばをていねいに一語一語切り取って数える方法しかない。しかし、この「一語」というのがなかなか判断のつきかねる場合が多い。

「国立国語研究所国語辞典編集室」を例として考えるならば、これを１語とみなすこともできる一方で、２語あるいはそれ以上とすることもできる。この、何語とみなすか、という点がつまりは単位というものを考えていることに他ならないのであるが、これを他の何万と出現する語との関係で一貫性を持たせていかなければならない。

国語辞典編集室でのその点に関する考え方について作業の流れも示しながら述べていくことにする。

２．用例データベース作成作業の流れ：文学作品

２．１ 対象とする文学作品

文学作品は、国語辞典編集準備資料２「用例採集のための主要文学作品目録」に挙げられて

いる、１９０１年から１９５０年までに発表された１１６点の作品を当初の対象とし、国定読本に対するのと同じ全数方式によって採集を始めた。これらの作品は、現代の代表的な文学全集１５種類に納められた文学作品のうち、３つ以上の文学全集に収録されている作品１５０６点を選定した後、１０名の所外専門家によってその１５０６作品から取り上げるべき文学作品を推薦して頂き、４名以上の方から推薦されたものである。その後、「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「用例採集のための主要文学作品目録」に挙げられている「主要文学全集収録作品目録」から、やはり概算で１万文節以下の作品を優先して調査対象としていくこととした。

２．２ 対象文学作品数

「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もっとも短い作品
合　　計	３３４作品	

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

２．３ 対象文学作品の種類と作品数

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

２．４ 底本とするもの

対象としている文学作品は、書誌調査を行い、初版を底本としてそのコピーあるいは復刻版を入手して用いるが、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若干残っている。

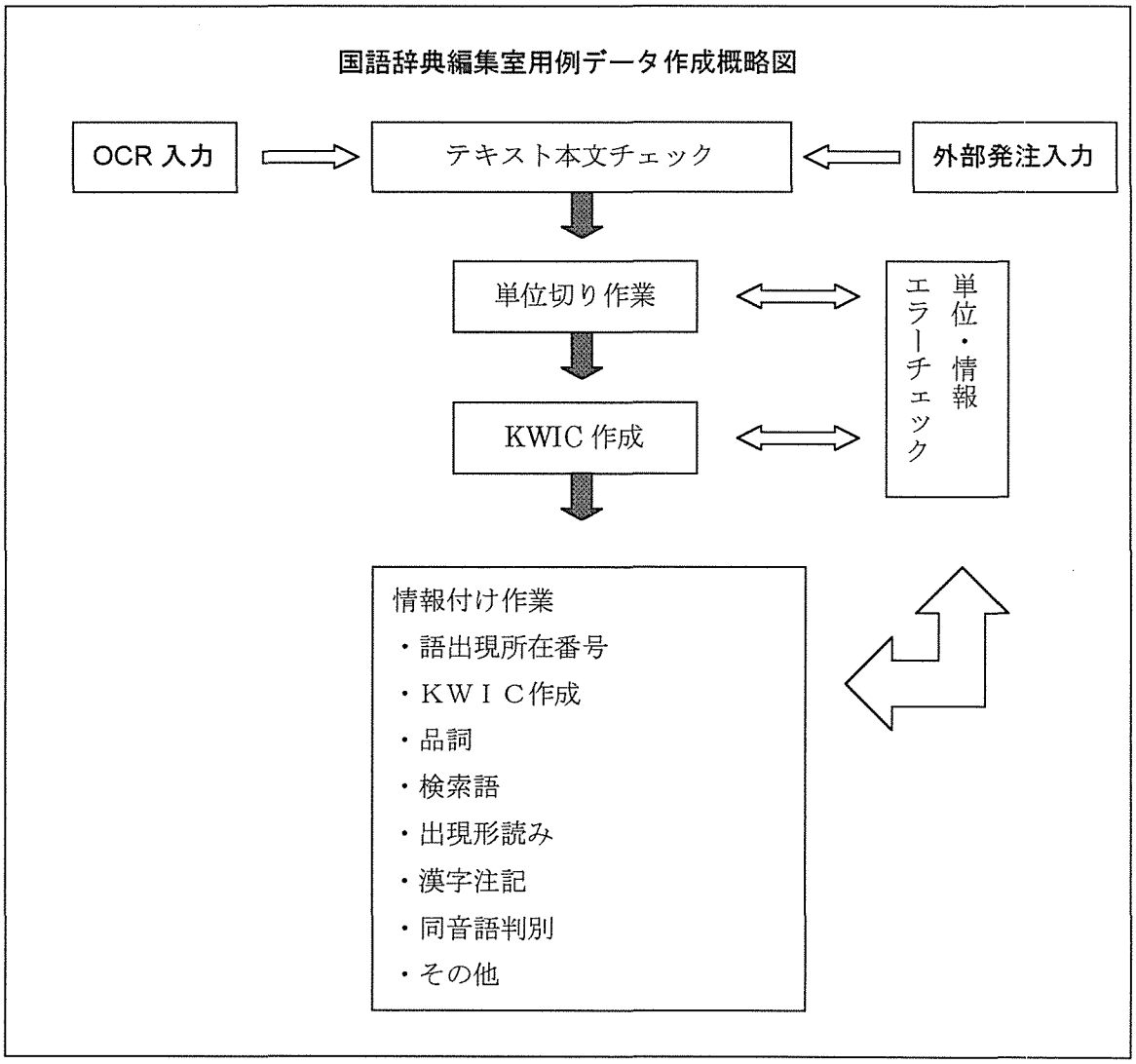
種　　類	作 品 数
初出単行本不明	１６
初出雑誌不明	１１
雑誌掲載なし	９
合　　計	３６

２．５ 著作権者調査

1998 年 12 月 1 日現在

説　　明	人 数
没後５０年以上経過作家	９９
著作権継承者判明作家	７０
小　　計	１６９
著作権継承者不明作家	3
存命中作家	3
小　　計	6
合　　計	１７５

３．新たな作業方法



近年、計算機の処理能力が向上し、また、周辺記憶装置の大容量化などもあって、大量の言語資料を計算機上で簡単に扱えるようになった。これにより、どの用例採集対象資料にも一貫した方針に基づいて採集を行う方法が可能となってきた。作業対象によって単位や採集方法を変えるというような従来の方法ではなく、単位の規準、旧漢字の処理規準などの統一化を図り、

データ自体も統一的なものとする新たな作業方法について述べる。作業の流れは、以下の通りである。

・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）
・加工されたテキストを短い単位に分割（単位切り作業）
・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なる検索語、読み、同音語判別などの情報を付与

これにともなって、計算機上でテキストのさまざまな加工段階に応じた共通作業マニュアル、テキスト上の旧漢字などをどう処理すべきかといった規準、見出し語として採用すべき語の単位についての規準などのマニュアルが作成される。（現在、実データを用いてＪＩＳコード上にない漢字の扱いや、新たに試行中の短い単位の有効性などの検討を行っているところである。）

３．１単位と検索語

ここに述べる「検索語」というのは、国定読本用語総覧などで見出し語、後要素などとして扱われていたものに代わる概念である。

前述の通り、テキストをある規準で単位に分割する単位切り作業では、できる限り短い単位で切ることを考えている。この分割されて切り出された単位が検索語のもととなる成分である。検索語というのは、単位そのもの、およびそれをいくつか結合させたものである。検索語ということをアイデアとして出してきたのは、「国立国語研究所国語辞典編集室」という文字列について何通りもの単位の切り方があることへの一つの解決策としてである。例えば、「国立国語研究所」はそのままか、「国立/国語研究所」、「国立/国語/研究所」として、一見出しから三見出しのどれかとして扱われることになる。「国立」の「立」、「研究所」の「所」などは、「公立、私立」、「裁判所、市役所」などとの関係で、「立」、「所」を後要素（主として造語能力のある共通成分）としてたて、「立」という後要素からも「国立」という見出し語にたどりつけるように工夫してある。ただ、こうした後要素を盛り込める形をとることができたのは、国定読本各期（全六期）それぞれの総体を見渡せる量だったからである。これが語数のさらに多い雑誌『太陽』や文学作品であれば後要素をもれなく盛り込む作業をするためには、そのつど作業を停止して、全体を注意深く見渡さなければならない。また、見出し語をたてるときに、その長さを決めるのにどうしてもゆれる場合がある。そのような作業をそのつど中断することなく、後要素や見出し語相当の情報を付与するような方法として検索語を設けることにしたのである。

国定読本用語総覧では他の語の長さとの関係で「戦艦ニコライ一世以下四隻」「巡洋艦以下數隻」のような場合に一見出しとして扱っており、後要素として「以下」から検索はできる、これらがそのまま「以下」の用例とはなっていない。こうした場合に、「戦艦」でも「艦」でも、また「ニコライ」であれ、「以下」や「隻」という語からでもそれらの用例を見ることができるようにしたのが検索語を用いたデータ作成方法である。

３．１．１単位切り結果

ここに示すのは、テキストに単位切り記号「/」によって単位切りを行った状態のテキストである。

/公/法/上/の/契/約/　/#
公/法/上/の/契/約/の/あり/や/なし/や/は/、/久しく/學/者/間/の/問/題/なり/し/#
が/、/今/や/之/【/こ/れ/】/が/存/在/を/認むる/こと/一/般/法/律/社/會/の/趨/勢/と/なれ/り/。/#
＄然れ/＄ども/余/の/如き/は/今/尚/其/【/そ/の/】/存/在/を/疑ふ/もの/なり/。/#
積/極/説/に/曰/く/、/國/家/が/法/規/を/以/て/自/己/を/制/限/する/は/、/國/家/が/#
本/來/絶/對/無/限/の/權/力/の/主/体/なり/と/いふ/論/據/より/、/當/然/＄來る/處/#
の/論/結/なり/。

３．１．２単位切り結果のＫＷＩＣデータ

これを、それぞれの単位をキーとしてＫＷＩＣ作成したものが次のページに示すデータである。

法を用ゐざるかを怪む。　公　＼法＼　上の契約　公法上の契約のありや,法,,,,
東京商業會議所の前身なる商　＼法＼　會議所若くは東京商工會は、政府,法,,,,
らず。唯其【その】第三者が　＼法＼　律上權力を有することを要するや,法,,,,
を怪む。　公法上の契約　公　＼法＼　上の契約のありやなしやは、久し,法,,,,
幸ひに東京商業會議所の選擧　＼法＼　は、今回より單記無記名の制に改,法,,,,

３．１．３検索語付与データ

さらにこれに検索語を付与すると以下のようなデータになる。

ゐざるかを怪む。　＼公＼　法上の契約　公法上の契約のありやな,公,公法,公法上,,
ざるかを怪む。　公　＼法＼　上の契約　公法上の契約のありやなし,法,公法,公法上,,
るかを怪む。　公法　＼上＼　の契約　公法上の契約のありやなしや,上,公法上,,,

３．１．４検索語付データ形式

データの形式
ＫＷＩＣデータ,出現単位,検索語１,検索語２,検索語３…,検索語ｎ,その他の情報,

このデータ形式を採用することによって、この方法の利点は、見出し語をどうたてるか、というゆれを吸収することができる点、検索語をたてることでテキストデータの用例をさまざま

な切り口で抽出できる点である。

３．１．５この方式の目指すところと問題点

ここでは、品詞付与については触れなかったが、検索語同士での同音語判別などで品詞を用いることを考えているが、積極的にすべての検索語に品詞を付与することは煩瑣な作業となることが予想されるため、現在のところは考えていない。よって、テキスト全体に含まれる品詞別の統計などをこの方式ではとりにくいといえる。この方式が重きをおいているのは、いうなれば、テキストデータの中にある語あるいはそれに準ずるような単位に対する検索性の高さであるといえる。

４．用例データベースを用いてなにができるか

では、これまで述べてきたような用例データベースが作成されると、そのデータからどのようなことを導き出すことができるかについて述べる。

ここでは、平成９年に出された「国定読本用語総覧 CD-ROM 版」を使ってそこから抽出される具体的なデータの例を見ていくことにする。

例１：国定読本用語総覧に現れた文語と口語の比較

国定読本用語総覧

検索条件 | 全件表示 | 環境設定 | KWIC表示 | ファイル出力 | 件数: 4 | 終了

見出し番号	見出し	漢字注記	品詞	同	1期	2期	3期	4期	5期	6期	合計
159230	すくない	少	形		2	9	12	7	8	7	45
159250	すくなし	少	形		6	39	13	3	0	0	61
215710	ない	無	形		48	124	224	308	287	407	1398
218420	なし	無	形		16	182	52	44	42	7	343

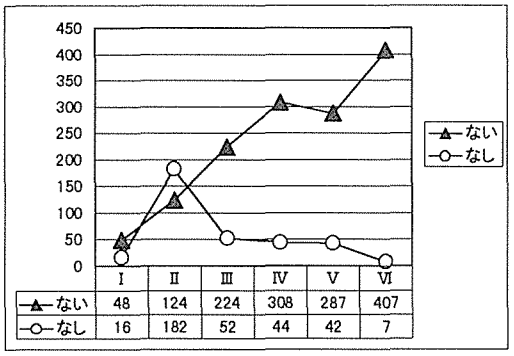
KWIC表示

159230, , すくない, 少, 80, 形, , 2, 9, 12, 7, 8, 7
159230, 13102306000, , ノヒカリヲ見ルコトガ, スクナク, 、夜モ、月ノ
159230, 13102306038, , ノヒカリヲ見ルコトガ, スクナイ, 。ニハニハ、
159230, 22203104010, 会, 、 は ねばりけ が , すくない, から、おも
159230, 22205401028, 会, 。オレ ノ 方 ガ , 少イ, カ モ 知レナ
159230, 24102606038, , フノハ、ハタラク人ノ, 少イ, トイフコトデス。
159230, 24102803034, , マセン。コレハチエガ, 少イ, カラデス。手バカ
159230, 24201905006, , り、畑の取高も年々に, 少く, なつて、五六年の
159230, 24202005026, , いて、近年麥の取高の, 少い, のは、この雀のせ
159230, 25208404000, , 牛肉を食ふ人は至つて, 少かつ, たが、今では全
159230, 26111007008, , とも、外出することも, 少い, 。京城地方の婦人
159230, 26202109044, , 酸瓦斯の分量は至つて, 少い, 。外に同化作用と
159230, 32201308036, 会, 、 タチ ノ 方 ガ , 少イ, カ モ シレナ
159230, 35101603040, , 敵ニオソハレル心配モ, 少ク, 、又コチラカラ敵
159230, 35102302024, 会, 、全く手の着かない事が, 少く, なかつた。そこで
159230, 35204605042, , つた。研究の爲には、, 少から, ぬ費用もかゝる
159230, 35208703022, , 買つた方が得な場合が, 少く, ない。それで、機
159230, 35208801016, , に外國へ輸出する事も, 少く, ない。綿花は主に
159230, 36103806000, , て、こずゑの差が段々, 少く, なつて行くのも面
159230, 36106305038, , 長々と續いてゐるのが, 少く, ない。こんな廣い
159230, 36109304042, 会, 、一回歸年より約十一日, 少い, から、太陽暦とく

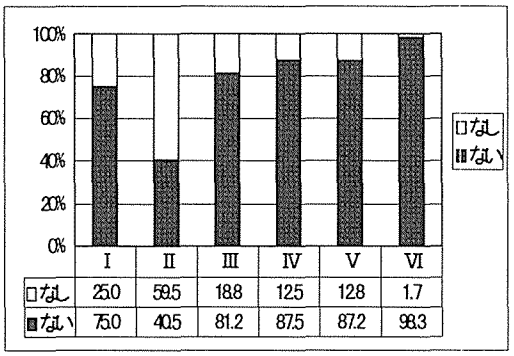
折り返し表示 | 閉じる

上の表は CD - ROM を用いて一期から六期までの出現数を検索によって得たところ、下の表はその検索結果のうちの「すくない」の KWIC を表示させたところ（部分）である。

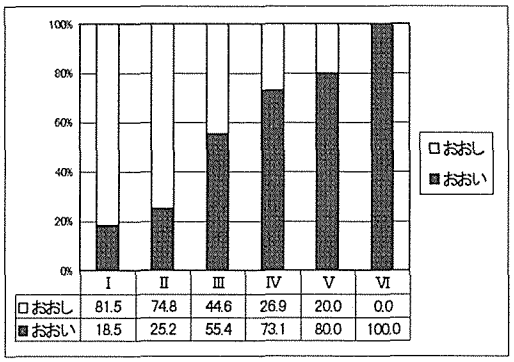
国定読本第一期（明治三十七年）から第六期（昭和二十二年）の六期の各期それぞれを単体の共時的なデータの集合とみなし、それを六期分並べることで通時的変化をデータから読み取ってみる。ここでは、文語「なし」「おおし」「すくなし」と、口語「ない」「おおい」「すくない」との出現のしかたを見た。Ⅰ～Ⅵは第一期から第六期を意味する。また、「なし」と「ない」の折れ線グラフは出現数そのものを比較している。その他の棒グラフは、文語・口語双方の各期での和を百としたときに、それぞれがどの程度の割合を占めるか、というパーセンテージに基づいて示したグラフである。



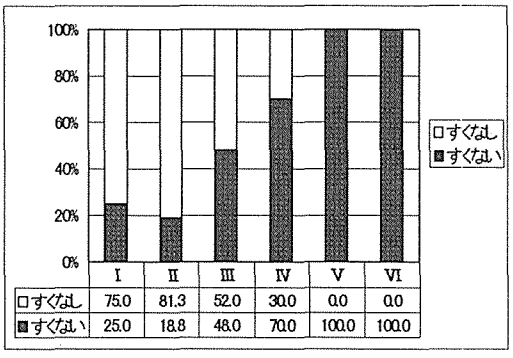
	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ
ない	48	124	224	308	287	407
なし	16	182	52	44	42	7



	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ
なし	25.0	59.5	18.8	12.5	12.8	1.7
ない	75.0	40.5	81.2	87.5	87.2	98.3



	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ
おおし	81.5	74.8	44.6	26.9	20.0	0.0
おおい	18.5	25.2	55.4	73.1	80.0	100.0



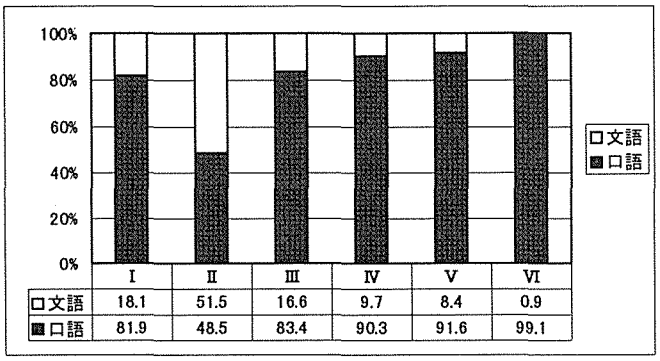
	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ
すくなし	75.0	81.3	52.0	30.0	0.0	0.0
すくない	25.0	18.8	48.0	70.0	100.0	100.0

次に、上記グラフと比較する意味で、国定読本用語総覧の層別情報を用いて文語・口語の出現数を各期毎に通観してみる。

	1期	2期	3期	4期	5期	6期	計
無印	16517	29105	54455	83589	87679	87312	358657
会	8338	6398	16143	18160	18112	31882	99033
韻	944	1000	1352	2987	3665	2243	12191
韻会	55	94	90	156	454	189	1038
手	699	964	4624	5585	5416	4263	21551
手会	0	0	95	28	221	97	441
文	4521	31829	10628	8101	5640	75	60794
文会	761	2458	915	1544	2507	0	8185
文韻	386	2654	1886	1770	2020	976	9692
文韻会	0	218	92	115	56	25	506
文手	0	86	0	0	60	4	150
文手会	0	0	27	0	0	6	33
候手	192	2554	1719	256	283	0	5004

※上記　無印：口語（地の文）　　手：手紙文　　会：会話文　　韻：韻文
　　文：文語（地の文）　　候：候文　　をそれぞれ意味する

では、これを「文」「候」の印のついたものを「文語」とし、それ以外を「口語」と考え、各期の総語数を１００として文語・口語各々のパーセンテージを求めてグラフにしてみたものが、以下のグラフである。



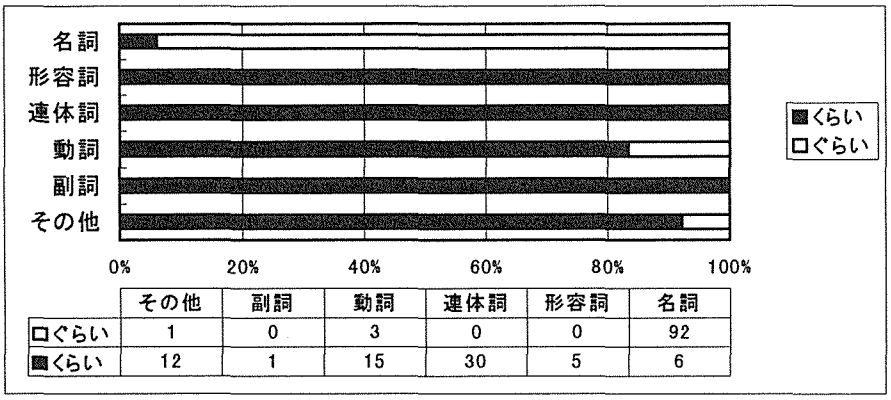
	I	II	III	IV	V	VI
口文語	18.1	51.5	16.6	9.7	8.4	0.9
口語	81.9	48.5	83.4	90.3	91.6	99.1

例２：国定読本に現れた副助詞「くらい」と「ぐらい」

では、次に、こうしたある程度の規模の用例を集めていえる例を挙げる。この例の場合、国定読本の第一期から第六期までを 1901 年から 1950 年までの間のまとまった共時的なデータと捉えて扱うということになる。

我々は「100 円くらい」、「100 円ぐらい」のどちらを我々は目にする機会が多いのだろうか。

副助詞「くらい」と「ぐらい」に使い分けがあるだろうか、国定読本では以下のような結果になっている。



	その他	副詞	動詞	連体詞	形容詞	名詞
ぐらい	1	0	3	0	0	92
くらい	12	1	15	30	5	6

このグラフは、「くらい」「ぐらい」の前にどのような品詞が来るのか、を「くらい」「ぐらい」の KWIC データの検索によって得た結果に基づいている。これを見る限り、国定読本では名詞については圧倒的に「ぐらい」を用いていることがわかる。

参考文献

国立国語研究所：国語辞典編集資料１～１２「国定読本用語総覧１～１２」

国立国語研究所：「国定読本用語総覧 CD-ROM 版」（市販品発行所三省堂）

加藤 安彦（1996）「国定読本における副助詞『くらい』と『ぐらい』」国立国語研究所研究報告集１７