語彙調査自動化のための基礎的研究

語彙調査自動化のための基礎的研究

中野　洋（言語体系研究部第２研究室）

１．はじめに

本発表は，特別研究『語彙調査自動化のための基礎的研究』（昭和59年度～63年度）の報告である。これまでに国立国語研究所は，電子計算機を用いて次の大量語彙調査を行なってきた。

	調査対象	成果発表年	標本数　ＯＨＰ①
(1)電子計算機による新聞の語彙調査	昭和41年新聞3紙 1年間	S45, S46, S47, S48	約300万β単位
(2)高校教科書の語彙調査	昭和49年理科社会科9冊	S58, S59, S60, H1	約 60万M単位
(3)中学校教科書の語彙調査	昭和54年理科社会科7冊	S61, S62, H1	約 25万M単位
(4)テレビ放送の用語調査	平成元年度 7CH 1年間		約 70万長単位

これらの調査においてきめの細かい分析を行なうためには，人手の作業がかなり必要となっている。予算および人員が削減される現状において，データの精度を維持するためには，作業の自動化を推し進め，人手と計算機が緊密に結びついたシステムを作り上げることが望まれる。

このことを実現するために，次の機能を持った自動処理プログラムとデータベースマネージメントシステムを試作した。

(1)修正しやすいこと，(2)検査しやすいこと，(3)結果が得やすい，(4)手軽なこと

２．語彙調査の自動化のためのプログラム

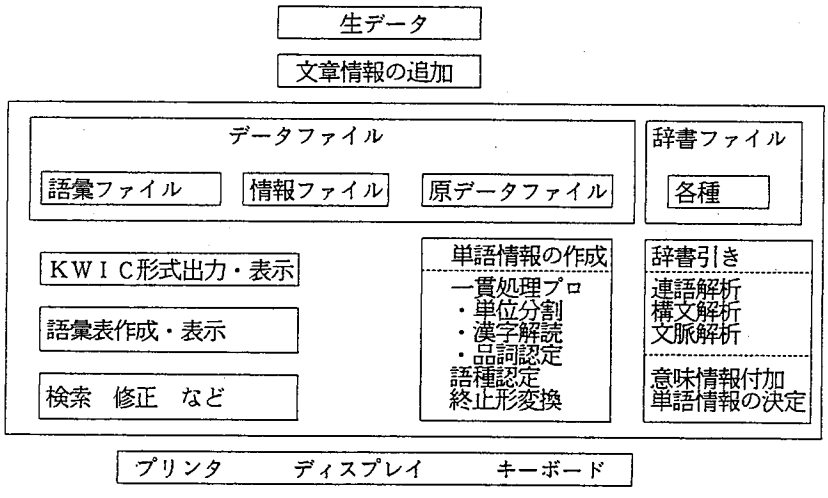
はじめに，本研究は，語彙調査を完全に自動化することをめざしているのではないことを断っておく。そのねらいは，自動化できるところを自動化し，その分人手による精度の向上を図ることである。

(1)語彙調査システムの歴史

ＯＨＰ②

世代	第１世代	第２世代	第３世代
対象	新聞３紙	教科書	テレビ放送
特徴	人手作業重視 前処理 処理量重視 電子計算機	人手・機械分離 前処理，後処理 精度重視 漢字プリンタ	人手・機械融合 後処理（ＤＢＭＳ管理） 精度・処理速度 大容量ディスク ＴＳＳ，パソコン

(2)語彙調査自動化システム



ＯＨＰ③

(3)データファイルの持ち方

(位置)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
原データ	新	し	い	計	算	機	シ	ス	テ	ム	と	の	関	連	性

OHP④

語彙ファイル	位置	読み	代表形	語種	品詞	意味番号
	1	あたらしい	アタラシイ	和語	形容詞	3.166
	4	計算機	ケイサンキ	漢語	名詞	1.4630
	7	システム	システム	外来語	名詞	1.132
	11	と	ト	和語	助詞	
	12	の	ノ	和語	助詞	
	13	関連性	カンレンセイ	漢語	名詞	1.1110

3. 一貫処理

(3.1)一貫処理の特徴①漢字かなまじり分を入力データとして，単語分割・読み仮名付け・品詞認定を行なう。②精度は９０％以上を目指す。③実用化できるよう，頑丈なシステムとする。④処理の精度より，処理速度を重視する。⑤小さい機械でも動くように，プログラムや辞書を小さくする。

(3.2)漢字解読の方法　漢字の前後の文字の種類によって，その漢字の読みを選ぶ。

OHP⑤

環境 直前	環境 直後	演算用コード A	1	B	2	C	3	D	4	E	5	F	6	G	7	H	8
非漢字	非漢字	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
非漢字	漢字	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1
漢字	非漢字	1	0	1	0	0	1	0	1	1	0	1	0	0	1	0	1
漢字	漢字	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1

０：漢字をテーブルの中の読みに代える

表1　環境演算テーブル

(1)　1校コウ　☆
(2)　2歌1カ　Aうた　☆
(3)　2河1か　Aかわ　☆
(4)　3川18セン　2Hかわ　＊M河1M柳1☆
(5)　3泳11エイ　2Aおよ　＊M平2Nぎ2☆
(6)　3水11スイ　2Aみず　＊M大2み気2☆

図1　漢字解読辞書

実験文1		実験文2		実験文3	
①	②	①	②	①	②
校	コウ	川	かわ	河	カ
歌	カ	で		川	セン
を		泳	およ	で	
歌	うた	ぐ		水	スイ
う		。		泳	エイ
。				をする	
				。	

①入力文字列

②漢字の読み

図2　漢字解読実験結果

(3.3)単語分割の方法　字種の変り目により切る。ひらがな辞書を引く。助詞・助動詞の接続を調べる。

漢字	ひらがな	カタカナ	数字	記
43．4%	28．0%	8.1	9.8%	9.2

0.6%　(ローマ字)

図3　新聞における文字の使用分布　　延べ字数　1,489,175

日本語文を文字の連続とみて，入力文を次のように字種の列に変えることができる。

ＡＭ．１０にバスに乗る。
英英記数数平片片平漢平記

OHP⑥

前＼後	漢字	平仮名	片仮名	英文字	数字	記号
漢字	5.7	61.7	45.2	75.0	100.0	73.8
平仮名	92.1	40.8	95.7	100.0	100.0	95.1
片仮名	25.4	89.5	1.0	—	—	33.3
英文字	2.8	100.0	100.0	13.2	0.0	90.0
数字	2.7	100.0	—	100.0	0.0	75.0
記号	98.2	84.7	62.1	33.3	23.7	—

表２．語の切れ目における文字種連続の割合

前＼後	漢	平	片	英	数	記
漢字	0	0	0	1	1	1
平仮名	1	0	1	1	1	1
片仮名	0	1	0	0	0	0
英文字	0	1	1	0	0	1
数字	0	1	0	1	0	1
記号	1	1	1	0	0	0

0:分割しない
1:分割する

表３．文字連続による単語分割の表

字数　文字列（10字以内）	①②③①②③①②③
1が	1R
4こうした	2C　1E91P
1た	1P＋
1で	1O9
1の	1R
1れ	1P＃

①：単語の長さ
②：品詞
③：活用

図４．品詞認定・単語分割のためのテーブル例

OHP⑦

文	区切り	文	区切り	文	区切り	文	区切り	文	区切り	文	区切り
C		ン		に	1	ョ		領	1	1	
O		タ		あ		ン		だ		0	
L		ー		き	1	・		っ	1	0	
I		ホ		た	1	F		た	1	円	
N		ー		子		・		。		分	1
G		ル	1	供		ケ		パ		く	
8		で	1	ら	1	ネ		ン		だ	
0	1	開		が	1	デ		粉	1	さ	
が	1	催		帰		ィ	1	を	1	い	1
東		さ	1	っ	1	は	1	1		。	1
京	1	れ	1	て		偉		0			
の	1	た	1	い		大	1	0			
都		。	1	く	1	な	1	g	1		
市		遊		。	1	大		か	1		
セ		び	1	ジ		統		,	1		

区切り欄が 1の箇所で語が切れる。

図５．単位分割実験の結果

(3.4)品詞認定の方法　テーブルによる。語形による。助詞・助動詞の接続による。

１番目の方法は図４のテーブルによる方法である。

OHP⑧

２番目の方法は，語末の形による方法である。

その規則を適用した場合の精度をそれぞれの規則の後に（）付きで示した。

①もし語末の文字が漢字か，片仮名か英文字であれば，その単語は名詞である。(94.4%)
②もし語末の文字が「い」であれば，動詞の連用形か，形容詞の終止形または連体形である。(86.2%)
③もし語末の文字が「く」であれば，動詞の終止形または連体形か，形容詞の連用形である。(83.4%)
④もし語末の文字が「る」であれば，動詞の終止形である。(95.8%)
⑤もし語末の文字が「れ」であれば，動詞の仮定形か，指示代名詞か，助動詞である。(92.9%)
⑥もし語末の文字が「ろ」であれば，動詞の命令形か，名詞である。(63.3%)
⑦もし語末の２文字が「かっ」であれば，形容詞の未然形か，動詞の連用形である。(74.2%)
⑧もし語末の文字が「っ」であれば，動詞の連用形である。(79.6%)
⑨もし語末の２文字が「漢字＋平仮名」であれば，それは動詞である。(94.4%)
最後の文字の平仮名の母音が／ａ／であれば，その語の活用形は未然形または連用形である。
／ｉ／であれば，未然形または連用形，／ｕ／であれば，終止形または連体形，
／ｅ／であれば，仮定形または命令形，／ｏ／であれば，命令形である。
⑩もし語末の文字が数字であれば，それは数字であり，語末が記号であれば，記号である。

３番目の方法は語の接続のしかたを利用する方法である。

ﾌｫｰﾏｯﾄは次の通り。(@は区切り記号である。)

①語　②品詞　③この語の直前に用いることのできる助詞・助動詞
④この語の直前に用いることのできる品詞と活用形
⑤もし，直前の語が３・４と一致しなければ強制的に適用する品詞・活用形

①　②　　③　　④　⑤
を@R@#と#から#まで#の#だけ#ばかり#こそ#さえ#すら#のみ#など#ぐらい#　1　／1

図６．品詞接続テーブル

OHP⑨

①	②	③	④⑤	⑥⑦
祭	まつ			
り		1	E #	1
を		1	R	R
待	ま			
っ		1	E 9	E 9
て		1	R	R
い				
る		1	E +	E +
。		1	Y	Y

①入力文
②漢字解読の結果
③単位分割の結果
④方法１・２の品詞認定の結果
⑤活用形
⑥方法３の品詞認定の結果
⑦活用形

品詞コード
1:名詞　A:接続詞　B:感動詞
C:副詞　D:連体詞　E:動詞
M:形容詞　P:助動詞　Q:助動詞,助詞
R:助詞　Y:記号　X:数字
活用コード
8:未然形　9:連用形　#:未然形,連用形
H:終止形　I:連体形　+:終止形,連体形
Q:仮定形　R:命令形

図７．品詞認定実験の結果

(3.5)スーパーバイザ　①助詞・助動詞の接続をチェックする。　②１字で構成される単語をチェックする。③　動詞の連用形＋他の動詞をチェックする。　OHP⑩

文字	頻度	助詞助動詞の頻度	%	その他の語の頻度	%	文字	頻度	助詞助動詞の頻度	%	その他の語の頻度	%
の	38404	32588	84.9	2	0	は	16062	13324	83.0	0	0
い	23633	2	0.0	1305	5.5	た	15958	10569	66.2	1	0.0
し	22124	64	0.3	13138	59.4	る	15522	17	0.1	0	0
に	18962	17037	89.8	3	0.0	を	14710	14702	99.9	0	0
と	16383	10173	62.1	0	0	で	13515	8351	61.8	0	0

図８．文字の頻度とその１文字語の頻度

OHP⑪

①	②	③	⑥⑦	③	⑥⑦
沢	たく				
山	さん	1	1	1	1
の		1	R	1	R
木	き	1	1	1	1
を		1	R	1	R
た		1	P+	0	
ば		1	R	0	
ね		1	Q	0	
ら				1	E8
れ		1	P#	1	P#
ま					
せ		1	P#	1	P#
ん		1	P+	1	P+
で					
し		1	P9	1	P9
た		1	P+	1	P+
。		1	Y	1	Y

①	②	③	⑥⑦	③	⑥⑦
面	おも				
白	しろ				
く		1	M9E+	1	M9E+
て		1	R	1	R
遊	あそ				
び		1	E#	0	
過	す				
ぎ		1	E#	1	E#
た		1	P+	1	P+
。		1	Y	Y	

図９．スーパーバイザの結果

４．単語情報の作成

(4.1)一貫処理の結果は，前の例に示したように文字単位で各種の情報を出力している。これを単語毎の情報に直す。さらに，語種の認定，終止形変換も行なう。辞書引きによる意味情報の付与機能はまだ付けていない。

(4.2)語種認定

語種の認定は，漢字解読テーブルの読み情報を利用する。漢字解読テーブルの読み情報は，訓読みは平仮名，音読みは片仮名表記になっている。外来語読みはローマ字表記となっている。

漢字表記の語はこれらの情報を利用する。仮名表記の語は，片仮名なら外来語，平仮名なら和語とする。

(4.3)終止形変換

文章中に現れた各活用形を終止形に変換する。同語異語の判別を助けるためのプログラムである。前後の文脈を調べないで終止形に変換するには，次の３つの方法がある。

処理方法	処理速度	辞書の大きさ	ﾌﾟﾛｸﾞﾗﾑ
①活用語辞書とのマッチング	遅い	大きい	簡単
②活用情報による終止形変換	早い	小さい	複雑
③出現形の漢字表記の利用	遅い	大きい	簡単

処理方法

入力データには，品詞認定の結果として活用形の情報が付いている。これを利用して次の処理を行う。

①形容詞・助動詞は，プログラム内の活用表によって終止形に変換する。　　OHP⑫

②動詞は，以下の方法による。

カ変・サ変は，プログラム内の活用表によって終止形に変換する。

終止・連体形は，そのまま出力する。

仮定・命令形は，語末の「れ・ろ・よ」を「る」に変える。それ以外は，　語末をウ段に変える。

未然形は，語末がエ段またはイ段なら「る」を加える。その他はウ段に変える。

連用形は，語末がエ段なら「る」を加える。イ段または促音・撥音ならテーブルにしたがって変換する。たとえば，「いった」はテーブルにしたがい，すべて「いく」と変換する。テーブルの内容は確率的に多い方を採用しておく。

5．処理の精度　以下は，大型計算機による結果である。パソコンによる処理の結果もほぼ同じである。

(5.1)結果のまとめ　　OHP⑬

1. 語彙調査データの作成作業における人手の作業と機械処理の比較を行った。
2. 処理精度は単位切りでは機械でほぼ90%，人手では97%～98%が見込まれる。これは人手の方がよい。
3. 処理時間は検査の時間を含めても機械が約5時間，人手が約53時間で，人手は機械処理の10倍かかる。
4. 入力パンチ量については，機械は人手の約20%の入力で済む。
5. 以上の結果，今後の語彙調査には機械による自動処理を用いても良いことは明らかである。しかし，今まで以上によい修正システムをつくる必要があると思われる。

(5.2)調査対象と機械処理の精度

調査対象　　OHP⑭

分類	対象	総字数	漢字%	手作業者
高校教科書	世界史	2548	40.6%	大学２年生
	政治経済	2067	37.2	――
	物理	2353	30.8	――
	生物	2642	33.3	――
雑誌	A. 中央公論	5430	42.7	教育学部卒
	B. 現代の眼	4787	31.5	――
	C. 主婦と生活	4947	24.5	大卒(1ﾃﾞｰﾀ

機械処理の精度

	単位切り	漢字解読	品詞認定
教科書	90.6%	90.1%	96.9%
雑誌 A	93.1	89.0	96.7
雑誌 B	89.7	92.5	95.6
雑誌 C	88.0	87.2	95.0

(5.3)人手と機械の精度について　　人手作業と機械処理の精度を比較したものが，次の表である。

	世界史			雑誌	
	機械	修正後	人手	機械	人手
単位切り	92.7	97.0	97.0	91.4	99.2
よみがな	92.2	99.9	100	87.9	99.7
品詞認定	97.3	93.5	91.7	96.0	99.9

(5.4)処理時間について　作業における処理時間をまとめたものが，次の表である。　　OHP⑮

	機械処理				人手作業				
	一貫処理	全体*	検査	単位	清書	かな	品詞	全体**	検査
世界史	０．１秒	３０分	４時間	２時	５時	２時	６時	３日	１時間
雑誌ﾃﾞｰﾀ	０．６秒	６４分	――	２時間	１１時間	７時間	９時間	６日	１時間

* LOGIN, LOGOUT, ｵﾍﾟﾚｰﾄﾐｽなどのすべてを含む

* 仕事につく前の時間・休憩なども，すべて含む

(5.5)パンチ量について　　パンチ量を比較したものが，次の表である。

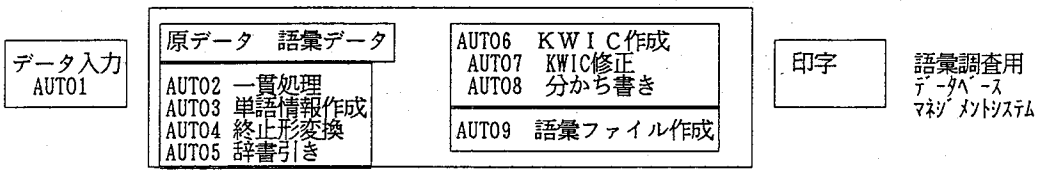
	語数	機械	人手	割合 機／人
世界史	1,296	2,548字	11,926字	21.4%
雑誌A	4,217	7,596字	37,531字	20.2

6．利用の方法　パソコンによる利用について示す。

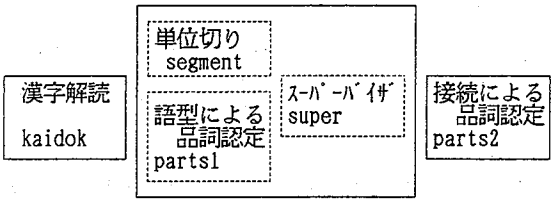
(6.1)動作環境

ＭＳ－ＤＯＳが動き漢字の使えるコンピュータであれば使用できるはずである。メモリーは空き領域１２８Ｋ以上あればよい。補助記憶装置は，ﾌﾛｯﾋﾟｰﾃﾞｨｽｸ1台でも動作可能である。しかし，出力ファイルが入力ファイルの８倍の大きさになり，又，中間ファイルも出力ファイルと同じだけの領域を必要とする。大量データを処理する場合は，ﾌﾛｯﾋﾟｰﾃﾞｨｽｸ２台，できれば固定ディスク装置があったほうがよい。

(6.2)システム構成



ＯＨＰ⑯



一貫処理プログラム

(6.3)辞書一覧

下の表に，一貫処理で使用した主要な辞書の一覧を示す。これらは，MS-DOSのテキストファイルで書かれている。他の日本語処理の辞書に比べ，大変小さいことが特徴となっている。また，辞書は，処理対象の文章に応じて書き換えることができ，処理の精度を上げることが出来る。

ＯＨＰ⑱

	辞書名	ファイル名	ﾊﾞｲﾄ数	項目数
①	漢字解読用テーブル	KAN.TBL	94477	2945
②	漢字解読テーブル用索引	KAN.IDX	17672	
③	単位切りテーブル(漢字仮名混じり用)	SEGMENT.TBL	10321	359
④	単位切りテーブル(仮名分かち書き用)	SEGREV.TBL	10327	359
⑤	助詞，助動詞接続テーブル	POSTBL1.TBL	1357	34
⑥	品詞接続テーブル	POSTBL2.TBL	297	15
⑦	助詞，助動詞接続チェック用テーブル	PRTSTR.TBL	6391	142
⑧	連用形変換テーブル	RENYOU.TBL	103797	3660

(6.4)プログラム一覧　一貫処理プログラム　語彙調査用ﾃﾞｰﾀﾍﾞｰｽﾏﾈｼﾞﾒﾝﾄｼｽﾃﾑ　ＯＨＰ⑰

	ﾌﾟﾛｸﾞﾗﾑ名	内容
①	KAIDOK.EXE	漢字解読
②	SUPER.EXE	単位切り，品詞認定１，ｽｰﾊﾟｰﾊﾞｲｻﾞ
③	PARTS2.EXE	品詞認定２
④	OUTPUT.EXE	清書出力
⑤	NAP.EXE	一貫処理ﾄﾞﾗｲﾊﾞ
⑥	NAPOUT.EXE	一貫処理ﾄﾞﾗｲﾊﾞ清書出力付
⑦	KAIDOK0.EXE	漢字解読　分かち書き用
⑧	PARTS.EXE	品詞認定　分かち書き用
⑨	NAP0.EXE	分かち書き用一貫処理ﾄﾞﾗｲﾊﾞ
⑩	NAPKANA.EXE	仮名文節分かち書き用

ﾌﾟﾛｸﾞﾗﾑ名	内容
⑪AUTO1.EXE	データ記述ファイルの作成
⑫AUTO2.BAT	一貫処理
⑬SUTO3.EXE	単語情報の作成・語種認定
⑭AUTO4.EXE	終止形変換
⑮AUTO5.EXE	辞書引き（予定）
⑯AUTO6.EXE	KWIC作成
⑰AUTO7.EXE	KWIC表による単位切り修正
⑱AUTO8.EXE	分かち書きデータの作成
⑲AUTO9.EXE	語彙ファイルの作成
⑳汎用プロ	MIFES,OPT-TECH SORT,SED,NFP

7．処理データ例　（原データ：保存ファイル）

＜題名＞「日英語彙データの収集・比較と機械辞書の作成」　（中略）　ＯＨＰ⑲

Ⅰ．研究目的

異なる言語を比較すると，ある言語では一つの単語で表される概念が他の言語では連語や句でなければ表せないことがある。

また，一つ一つの単語では対応がとれるが，連語や句になると全く異なった表現となることがある。

（情報ファイル：保存ファイル）

```
  11  1  1「日英語彙データの収集・比較と機械辞書の作成」
 572  2  1001
 735  3  1    国立国語研究所・中野洋
1055  3  2    宮島達夫（国立国語研究所）
1415  3  3    石井久雄（国立国語研究所）
1775  3  4    藤田正春（国立教育研究所）
```

（単語ファイル：一時ファイル）

```
287 4,SE+,異,ことなる,異なる,ことなる
291 4,T1 ,言,げんご,言語,げんご
295 2,SR ,を,を,を,を
297 8,VE+,比,ひかくする,比較する,ひかくする
305 2,SR ,と,と,と,と
```

（ＫＷＩＣファイル：一時ファイル）

```
283 2      Ｉ．研究目的    異なる言語を比較,YY ,  ,  ,  ,
287 4    Ｉ．研究目的    異なる 言語を比較す,SE+,異,ことなる,異なる,ことなる
291 4 研究目的  異なる 言語 を比較すると、,T1 ,言,げんご,言語,げんご
295 2 目的  異なる言語 を 比較すると、ある,SR ,を,を,を,を
297 8 的  異なる言語を 比較する と、ある言,VE+,比,ひかくする,比較する,ひかくする
305 2 る言語を比較する と 、ある言語では一,SR ,と,と,と,と
```

（語彙ファイル：保存ファイル）　　ＯＨＰ⑳

```
が,(が),和語,助詞,[が]=  345 2,  389 2,  427 2,  435 2,  481 2,
く,(句),漢語,名詞,[句]=  365 2,  445 2,
げんご,(言),漢語,名詞,[言語]=  291 4,  313 4,  351 4,  511 4,
こと,(こ),和語,名詞,[こと]=  385 4,  477 4,
ことなる,(異),和語,動詞,[異なっ]=  287 4,  459 6,
ひかくする,(比),混種,動詞,[比較する]=  297 8,
```

参 考 文 献

1. 石井　正彦「自動単位分割の精度と問題点」(CL通信第3号,1986.4,10,国立国語研究所言語計量研究部)
2. 江川　清「漢字かな混り文の「自動単位分割」に関する一研究」（計量国語学第４３／４４巻，1968）
3. 江川　清「単位分割自動化のシステムについて」（計量国語学第５１巻，1969）
4. 小沼　悦「一貫処理プログラムの評価実験（３）――精度，人手作業との比較において――」（ＣＬ通信第３号，1986.4，国立国語研究所言語計量研究部）
5. 田中　章夫「漢字かなまじり文を全文カナ書き・ローマ字書きに変換するシステムについて」（電子計算機による国語研究Ⅱ，1969，秀英出版）
6. 中野　洋「品詞認定の自動化」（電子計算機による国語研究Ⅲ．1971，秀英出版）
7. －「言語研究における一貫処理の研究」（電子計算機による国語研究Ｘ，1978，秀英出版）
8. NAKANO Hiroshi,TSUTIYA Shin'iti, TURUOKA Akio「AN AUTOMATIC PROCESSING OF THE NATURAL LANGUAGE IN THE WORD COUNT SYSTEM」(Proceedings of The 8th Internatonal Conference on Computational Linguistics,1980)
9. －「ひらがなの使用頻度とひらがな一字で表記される語の頻度数」（季報1980－夏号，国立国語研究所言語計量研究部）
10. －「語彙調査の自動化における一貫処理システム」(CL通信第1号,1985,国立国語研究所言語計量研究部)
11. －「自動漢字解読の精度と問題点」（ＣＬ通信第３号，1986.4，国立国語研究所言語計量研究部）
12. －「自動品詞認定の精度と問題点」（ＣＬ通信第３号，1986.4，国国語研究所言語計量研究部）
13. 国立国語研究所「電子計算機による新聞の語彙調査」（国立国語研究所報告３７，秀英出版，1970）
14. 国立国語研究所「高校教科書の語彙調査」（国立国語研究所報告７６，秀英出版，1983）
15. 国立国語研究所「中学校教科書の語彙調査」（国立国語研究所報告８７，秀英出版，1986）