解　説

日本語分析資料およびツールの調査†

長尾　真††　辻井潤一††

1. まえがき

自然言語の計算機処理は数値計算や他の一般の計算機処理とはかなり異なったむつかしさがあり，その研究にたずさわる人々はこれまでずっと苦労を重ねて来た．たとえば単語の語尾処理のプログラムなども，いくつかの所でいろいろ作られては捨てられ，また類似のものが作られるということがくりかえされて来た．言語データについても同様である．過去の仕事の上へ積みあげつつ進歩をしてゆくということがかなりむつかしいのが自然言語に関する研究の特徴といえるかもしれない．

しかし現在のように自然言語処理が活発に研究されるようになってくればくるほど，これまでに作られたかなりの量と範囲におよぶ自然言語処理のためのプログラムとデータの財産を，お互いに利用しあい，その上にさらに研究成果をつみあげてゆくことが大切となる．

このような考え方から自然言語処理用のプログラムとデータについて公開できるものだけでも集めて一欄表の形にすることが有意義であるということになり，その調査を行った．調査はなるべく広い範囲を心がけて行ったが調査対象としてぬけているところも非常に多いと想像される．また調査依頼をしたが返事をもらえなかったところも多い．そういったことから集められた資料は非常に任意的で，完全という点からすれば全く不満足なものとなってしまった．これを学会誌に一覧表としてのせることは客観性と公平を欠き，あまり意味がないということも考えた．しかし，現にその研究にたずさわっている我々の場合でも案外他の場所でのこのような資料の存在を知らないことが多いし，これから自然言語処理の研究を行う人々のためにも，少しでもこの種のデータが整理されることが望ましい．さらに今後この種のプログラムやデータを作成する時の参考にもなり，その移植可能性に対する配慮への刺激にもなる．このような理由から，ここで集めた資料を今後のこの分野の健全な発展のための一つの刺激剤としてみていただくことに意義をみとめ，種々の批判があるだろうことを覚悟でのせることにした．どうかその主旨を了解いただいて，この調査の不備をお許しいただきたい．

集めた資料はここに載せたものよりさらに詳しいものであったが，紙面の都合上簡単な形で記載せざるをえなかった．詳しいことを知りたい方はそれぞれの責任者に問いあわせられたい．

2. プログラムに関する調査

日本語処理用のプログラムとして，現在開発されているものを，表-1 の種目に分類して調査を行った．各分類項目について，気のついたことを以下に示す．

1　Parser：文法を与えることによって，入力文の解析を行うプログラムである．集約されたプログラムのうちで，使用プログラム言語を LISP としたものが5件（拡張 LINGOL, ATNJ, PLATON, LINGOL-K, J-ANALYSER），PL/1 としたのが2件（AUTO-SEG II, ヤチマタの parser），FORTRAN が1件（文型表を用いた文構造解析プログラム），アセンブラが1件（CRBP）である．LISP を使用言語とするシステムは，一般的傾向として意味処理に重点を置いており，また，ルール中に procedurer を書き込める等の特徴を持つものが多い．これは，データとプログラムを区別しない LISP の特徴を反映したものであろう．これに対して，PL/1, FORTRAN を使用言語とするシステムには，大量の言語データを処理することを目指したものや実用システムの一つの構成要素になっているものが目立つ．

2　KWIC, concordance：国語研が，言語調査の過程で開発した各種の KWIC, concordance に代表されるように，言語研究・言語調査を計算機によってサポ

† Collection of Data and Programs for Japanese Processing by Makoto NAGAO and Jun-ichi TSUJII (Department of Electrical Engineering, Kyoto University).

†† 京都大学工学部電気第2教室

表-1 Program に関する調査

1. Parser
2. KWIC, その他の concordance プログラム
3. 日本文エディタ
4. 形態素解析プログラム
5. 言語処理応用システム
6. ユーティリティプログラム
7. その他

	プログラム名	組織		備考
1	AUTO SEG II	茨城大	石綿 敏雄	・国語研究所で開発された文解析用のプログラム．文法定義，文法のコンパイル，文解析の実行ができる．実行結果は並列的に出力される．
	拡張 LINGOL	電総研	田中・佐藤・元吉	・MIT の Pratt の開発した LINGOL に幾つかの機能を施したもの．拡張機能の主なものは，(i) 分ち書きを自動化するプログラムがパーザに組み込まれている，(ii) パージング過程で予測制御機構を動作させることができる，(iii) 新しいユーティリティプログラムが付加されている．
	ATNJ	東京理科大	溝口 文雄	・Woods の ATN モデルを基本にした文解析，文出力プログラムで，自然言語の意味解析により意味のネットワークを作成したり，意味のネットワークにより文を生成したりできる．
	汎用パーザ（ヤチマタのパーザ）	日本 IBM	藤崎哲之助	・BNF で記述された context-free 文法と，Lexicon に基づき，文の解析を行い，意味解釈のための木－構造を生成するパーザである．
	PLATON	京 大	辻井 潤一	・Woods の ATN parser を基本にして，入力文中の任意の個所を切り出せ，ルールを書く時に並列的な処理を指定できる 'パターン・マッチング機能' と，最も有望な解析パスを優先して適用してゆくために必要な '柔軟なバックトラッキング機能' が付け加えられている．
	LINGOL-K	京 大	堂下修司・西田豊明	・拡張 LINGOL から自動分ち書きの機能を省略したものであるが，パージングに意味情報をより柔軟に反映させられるように予測制御の機能に拡張を加えている．
	文型表を用いた文構造解析プログラム	防衛庁	島貫 隆光	・機能語分析プログラム出力の品詞語ファイルを読み込み，(1) 品詞語を1センテンス分読み込む，(2) 述語の抽出，(3) 複文パターンテーブルにより単文への分解と複文構造の指定，(4) 単文の文構造の解析，(5) 体言の修飾関係の解析，(6) ロール付けなどの処理を行う．
	J-ANALYSER（仮名）	電総研	池田 尚志	・ローマ字分ち書きの入力文からその意味構造表示を得る．同音異語・多義語の処理，省略・参照表現の同定などの分析も扱っている．'成り立ちの良さ' の概念を導入し，多様な解釈がある場合にはその '成り立ちの良さ' で順序づける．
	CRBP	大府大	西田富士夫	・事実関係を記述する基本的な構造をもつ文，文献標題や特許請求範囲文などの名詞句や名詞節を入力し，入力文をカテゴリ照合の手法を援用した変形順位文法により，bottom-up 方式で構文解析し，係り受け構造の明確な入力文の内部表現形式を出力する．
2	MCLKWC	国語研	斎藤 秀紀	・FBCDIK コードで作成された KWIC レコードを読み，OMR 80 ラン連続シート上に KWOC 形式で出力するプログラム．
	KKNMAT/KKNEDT	〃	〃	・KKNMAT: OMR シート上にプリントされたマークおよび記入済みマークを読み，対応する漢字かなまじり文へ挿入するプログラム． KKNEDT: KKNMAT で処理されたファイルを KWIC 形式に編集するプログラム．
	CNVKP	〃	〃	・KWIC イメージ（漢字モード）の入力レコードを読み込み，OMR 80 ランカードイメージで漢字プリンタに出力するプログラム．
	TNK 02～04	〃	田中 卓史	・80 ランカードにパンチされた入力文を KWIC にして出力する．
	索引作成プログラムライブラリ	〃	中野 洋	・カナ，英文，漢字かなまじり文で単位切りされたデータを入力し，KWIC，WORD COUNT LIST を出力するプログラム．
	日本語文体調査	東京農工大	福原満洲雄	・日本語カナ書き文をデータとし，指定された文字列を含む文を出力する．
	ALPS-Concordance プログラム	電総研	坂本 義行	・種々の言語で記述されたテキストから語を単位とした KWIC および用語統計を計算し，出力するプログラムである．
	カナ KWOC リスト作成プログラムシステム	日立製作所	三浦 武雄	・語彙辞書を作成するための語彙調査用のカナ KWOC リスト作成プログラムである．
	遺跡関係用語表作成プログラム	奈文研	田中 琢	・入力された日本語テキストから，字種および付属語に関する簡単な前処理を行い，用語（特に，その分野を特徴づける名詞）の抽出を行う．この抽出された用語のコード順ソートおよび出現頻度表を作る．
3	EDIT 2	東 大	荻野 綱男	・TSS 用漢字エディタで，コマンドとしては TOSBAC-5000 のテキストエディタに類似したものが入れられる．
	漢字テキスト・エディタとランオンシステム	電総研	坂本 義行	・TSS 環境下で，漢字データを含むテキスト・ストリングを編集し，ユーザの見易い形式にして出力することを目的とする．

	プログラム名	組織	備考
3	FDMS (和文エディタ)	富士通　大高　義宏	・TSS 用のエディタ．従来の TSS 用エディタの各種コマンドを漢字かな混りデータに対して適用できるようにしたもの．現在作成中．
	日本文エディタ KOED	東京農工大　小谷　義行	・タブレット・ディジタイザ（座標読み取り装置）の上での指示により，漢字かな混り文をエディットする．出力された文書をはりつけたディジタイザの部分を直接的にポイントすることにより行うエディット機能ももつ．
4	自動文節認定システム	電総研　坂本　義行	・日本語の漢字かな混りテキストを入力し，大量の辞書を用いず，字種およびひらがな列の処理により，分節分ち書きを行うプログラムである．
	ローマ字による漢字入力プログラム	〃　〃	・計算機へ直接日本語を入力する一方法として，TSS 画像端末から会話形式で，ローマ字により入力し，同音異字の判定は表示選択により行うプログラムである．
	日本語の活用処理ルーチン	〃　〃	・文節単位に分割されているデータに対してその右端より，左方向へ活用，接続処理を行うプログラムである．
	漢字列の自動分割プログラム	京　大　辻井　潤一	・長い漢字列を，1文字語（独立語，接頭語，接尾語），2文字語に分割する．
	漢字かな混り文自動分ち書きプログラム	〃　〃	・日本語「漢字かな混り文」から，各単語の品詞付けを行う．結果は文法的に可能なすべての品詞付けである．
	ACC 1	東京女子大　水谷　静夫	・日本語の活用する詞に始まる活用語尾および後続する辞の結合について，その合文法性をチェックするアクセプタである．
	文節解析プログラムと係り受け解析プログラム	電電公社　島津　明	・文節解析プログラムは，語彙辞書に基づき，日本語文（ローマ字/カタカナ）の文節を識別し，文節情報（形態情報）を解析する．係り受け解析プログラムは，文節解析プログラムにより解析された文節間の係り受け関係を係り受け規則あるいは格関係パタンにより求める．
	機能語分析プログラム	防衛庁　島貫　隆光	・文節作成プログラム出力の文節を読み込み，シソーラス照合によるキーワードの自動認定とその認定で得た情報を基にし，付属語表を参照する事により，付属語分析を行う．これにより文節構成語の認定を行う．
	文節作成プログラム	〃　〃	・ひらがな列から非ひらがな列に変わる点，記号類を文節の切れ目とし，これに従い，漢字かな混り日本語文を読み込み '文節' に区切るプログラムである．
	一貫処理システム (NAP)	国語研　中野　洋	・分ち書きされていない漢字かな混り文を入力とし，コード変換・自動単位分割・自動よみがなつけ・品詞認定・意味番号つけを行う．
5	文章カナ漢字変換プログラム	日立製作所　三浦　武雄	・カタカナ書き日本語文を読み込み漢字かな混り日本語文に変換するプログラム・システム．
	速記反訳システム	〃　〃	・速記用特殊タイプライタで入力した速記記号列を漢字かな混り日本語文に変換するプログラムである．
	宛名かな漢字変換システム	〃　〃	・顧客用の姓名辞書ファイルを使うことにより，宛名のかな漢字変換を行う．姓名辞書をメンテナンスするためのサブシステムも用意されている．
	カナ漢字変換処理プログラム	九州芸工大　稲永　紘之	・分ち書きされたカナ文字文を入力し，漢字かなまじり文で出力するプログラム．
	漢字点字による自動代筆システム	電総研　坂本　義行	・視覚障害者が点字タイプライタを打鍵することにより穿孔された紙テープと点字モニタを計算機に入力し，プログラムにより普通文字へ変換し，出力する自動代筆システムである．
6	英和辞書検索プログラム	電総研　坂本　義行	・機械翻訳を目的とした英和辞書の検索システムである．基本単語（2,000語）とこれに付加された，屈接語，派生語（8,000）語について，全品詞とその訳語情報も出力される．
	日本語文節辞書作成・検索プログラム	〃　〃	・特許公報・第12類金属の加工見出し2,000，プログラミング部門見出し600についての自立語辞書である．検索は文節単位に分ち書きされているテキストを用いる．
	漢字ソートプログラム	NEC　鵜木	・漢字ソートは，漢字データ群を，①音読順，②訓読順，③部首順，④画数順，⑤ユーザ指定順などの一定の基準に従って並べるものである．
	GLAPS	東　大　荻野　綱男	・方言調査分析用に作成された統計パッケージの一種である．言語地図，グロットグラム，クロス集計表，頻度分布表などが作れる．
	アジア，アフリカ諸言語	東外大　AA 研	・アジア，アフリカ諸言語の語句の用例集を作成することができる．言語ごとにパラメータをとりかえる必要がある．
7	SLIPP II	大分大　岡田　直之	・幾つかの要素的図形から構成される図形の系列が与えられると，その中で生じている変化のもつ意味内容を解釈し，その結果を自然語で記述する．
	Friedman Model	日本 IBM　藤崎哲之助	・N. Chomsky の Aspects 流に記述された文法，Lexicon に基づいて文をランダムに生成し，その文法が正しいかどうかを検討する "grammar tester" である．
	PRODUCTION SYSTEM	東京理科大　溝口　文雄	・時間要素を含んだ意味ネットワーク上での推論の処理を行う．PS は書き換えルール，意味のネットワークデータ，ルールを適用する際の制御の三つのサブシステムより成る．

	プログラム名	組織	備考
7	音声実時間処理に関するプログラム GS 1 (言語習得 system 1)	北大 伊福部 達 東京理科大 佐伯 胖	・単音節音声認識用プログラム，発声制御実験用プログラム，聴性誘発脳波検査用プログラム．音響処理用プログラムなどがある． ・J. R. Anderson の HAM の表現方法を絵で与え，それに対する主題および文を提供することにより，I. M. Schlesinger のような位置規則を習得し，文を生成する言語習得モデルのシミュレーションプログラムである．

ートするためのシステムが多く開発されている．しかしながら，奈良文化財研究所の開発しているプログラムのように，他の研究目的（たとえば，ある特定研究分野の用語調査）のために作成されているシステムも報告されている．この種のシステムは，今回のアンケートの対象とした研究グループ以外でも，かなり作られている可能性がある．また，漢字入出力が以前に比らべて，かなり楽に使えるようになった現在，この種の「言語処理技術」の他分野への応用が今後飛躍的に増えると思われる．

3 日本文エディタ：日本語を対象とするエディタということで，特に入力・出力に工夫がこらされているものが多い．今回のアンケートの結果では，データ・エディタ，すなわち，別途バッチ的に入力された言語データを，TSS モードで修正・校正するためのエディタが4件あった．英文の場合の Run-off のような，テキスト・エディタ的なものは回答には含まれなかった．これは，まだ一般の使用者が簡単に漢字かな混り文を入力し，個人用のドキュメント作成を計算機援助の形で行えるまでには至っていないためであろう．personal use の漢字入力装置，あるいは，かな漢字変換の技術等が進歩するに伴って，この種の文書作成用のテキスト・エディタが開発されてゆくものと思われる．

4 形態素処理プログラム：日本語は，英語における空白のようにはっきりとした単語間の区切記号がないため，2で示した各種コンコーダンス，用語調査を行う場合にも，まず単語の認定を行う必要がある．この部分の処理は，品詞・活用形等の比較的よく整理された情報の他に，日本語固有の字種情報，例外に対する処理等を含む必要がある．現在の段階では，自立語辞書を使用するかしないかで二つのグループにわかれるが，後者のプログラムでは 98% 程度の精度を達成しているものが多く，用途を限ると実用の段階に達しているものが多い．

また，この形態素処理はこれ以後の日本語処理の基礎となると同時に，かな漢字変換・漢字かな変換等のための基礎技術になっている．日本語固有で，しかも実用的な言語処理を行うために是非解決すべき問題に，長い漢字連続の処理があるが，この部分の技術は今後の研究によるところが多い（本調査では解答がなかったが，JICST の「語基」による手法がある）．

5 言語処理応用システム：主として4の形態素処理レベルでの技術を，実用システムに適用したもので，かな漢字変換的なものが多い．

6 日本語処理用ユーティリティプログラム：各種の統計，ソーティング，辞書検索のためのプログラム等を含む．

7 その他：研究的色彩の強いシステムを，このカテゴリーに分類した．意味処理 (SLIP, PRODUCTION SYSTEM, GS 1 etc)，音声情報処理，言語調査用の統計パッケージ等さまざまな目的のプログラムが作成され，報告されているが，こういった分野での調査はまた別途行う必要があろう．このリストは，回答されてきたものを羅列しただけで，網羅的なものではもちろんない．

3. データに関する調査

日本語処理用データとして計算機に蓄積されているものを，表-2 の種目に分類して調査を行った．各分類項目について，気のついたことを以下に示す．

1 テキスト・データ：言語の研究を行う上での基礎資料として，テキスト・データは重要である．テキスト・データには，調査目的に合わせて予め人手によって補助情報を付加しているもの（国研，IBS，京大(化学教科書)）と，テキストをそのまま入力しているものの2種類に分類できる．

2 辞書的データ：このカテゴリのデータの代表的なものは，国語辞典・英和辞典といった辞書である．現在，電総研と京大によって国語辞典（三省堂出版・新明解国語辞典）のデータベース化が実現されており，また文部省特定研究言語のグループによって，英和辞典（三省堂出版・新コンサイス英和辞典）の入力とデータベース化が進められている．この他，日本語の処理を目指して，機械処理用の各種言語データが準備されつつある．この種のデータは品詞類の設定等が

表-2 DATA に関する調査

1. テキスト
2. 辞書的データ
3. 日本語（あるいはその他の言語）に関する統計データ（ひらがな連続，文型パターン，単語の頻度）
4. 文法規則
5. その他

	データ名	組織		備考
1	「星の王子さま」6か国語版1～5章	国語研	中野　洋	・「星の王子さま」の仏，日，英，独，西，伯の6か国語の1～5章を原本のイメージどおり入力してある．日本語はカタカナである．
	アジア，アフリカ諸言語の言語資料	東外大	AA 研	・チベット，中国，タイ，クメール，朝鮮，ヒンディ，アラビア，スワヒリ語の現代小説・新聞等の文と満州の満文老檔．
	数学教科書	東京農工大	福原満洲雄	・旧制中学の数学教科書をカナ書きにして，ベタ打ちにしたもの．数式・図表などはかなり省略してある．
	ドイツ語文	茨城大	石綿　敏雄	・ヘルマン・ヘッセ著「青春は美し」の原文．
	漱石・鷗外の文学作品，新聞，教科書その他	国語研	斎賀　秀夫	・β単位，または C, L, S 単位に切られた漢字かな混り文で，各語に語種・品種・品詞・活用・よみがなの情報がついている．（1）新聞 約300万語（2）文学作品 約89万語　（3）高校教科書 約60万語　（4）分類語彙表 約3.5万語
	DOCTRINA, GOPASSION	東　大	豊島　正之	・ドチリナキリシタン（1592年　天草刊）約2,500行とスピリツアル修行（1607年　長崎刊）第2部御ばっしょんの観念約3,500行をほぼ原文のままローマ字で入力してある．
	日本語構文 tree，かっこ付きデータ KWIC	IBS	木村　睦子	・中学理科教科書，特許公報，刑事判例集などのテキストを2分割法により構文木を書き，木の枝をかっこに，自立語を品詞に置きかえ，助詞・助動詞はそのままでコーディングし，これを左かっこのみをキーとして KWIC 形式に出力した．
	特許請求範囲の文	東　芝	野寄　雅人	・日本特許分類 99 (5) Ho「固体半導体装置」の特許公告の請求範囲の文．昭和39年～51年．
	遺跡関係用語	奈文研	田中　琢	・世界考古学大系 1, 2, 3, 4巻の全文（4万4千字）．日本考古学辞典から考古学関係用語・関連する学術用語を取り出し単語ごとに区切って入力したもの．倭名類聚抄の地名索引．
	計算機マニュアル	京　大	辻井　潤一	・FACOM データセットユーティリティのマニュアル文（図および表はのぞく）について，おのおのの英文と日本語文とを対にして入力．機械翻訳の基礎資料．
2	雑誌九十種語彙・表記データ	国語研	中野　洋	・雑誌90種を対象とした語彙調査の結果データ．語彙と度数とその表記形のバラエティを含んでおり，語に大分類的な語種と品詞がついている．
	自然語および図形処理用データベース	大分大	岡田　直之	・図形の言語的解釈システム SLIPP II の機械辞書．具象的および抽象的線図形を付随する情報と併せて格納した図形部門，日および英語単文を合成するのに必要な辞書類の言語部門，および基本的な事象の意味を指定してある意味部門からなる．
	新明解国語辞典	京　大	辻井　潤一	・電総研で入力された新明解国語辞典を処理加工したもの．使用コードは JIS で，各辞書項目（品詞，活用型，重要語マーク，かな書き，漢字表記）の定型データの section と意味記述の section がある．
	接頭字・接尾字テーブル	〃	〃	・当用漢字の漢字列中，接頭字・接尾字として使われた頻度をカウントしたデータ．各漢字が接頭字として使われる傾向の強弱を表現している．長い漢字列を要素の単語に分割するのに有効．
	助詞・助動詞等付属語テーブル	〃	〃	・日本語処理で必要となる助詞，助動詞，活用語尾，ひらがな書き自立語，語幹にひらがなを含む自立語 etc のテーブル．漢字かな混り文を処理し，文中の語の品詞を決定するのに有効．
	新明解国語辞典	電総研	渕・横山	・三省堂新明解国語辞典の本文すべてを，SIS 標準文字コードのデータ列として格納したもの．
	姓辞書，姓字単位辞書，名辞書，法人名辞書，地名辞書	日立製作所	三浦　武雄	・姓，名，法人名のかな表記，漢字表記および出現頻度等が収録されている．地名辞書には，北海道，東京都，神奈川県，京都府の地名のかな表記，漢字表記および地名コードが収録されている．
	カナ漢字変換用辞書	九州芸工大	稲永　紘之	・岩波国語辞典の見出し語を基礎に，各種の辞典・新聞雑誌の記事を参考に追加登録した派生語，長単位の自立語，慣用句，係り受けの句からなる辞書で，現在登録件数約21万5千である．
	付属語辞書作成用テーブル	〃	〃	・名詞，用言，助動詞を活用語尾により15に分類し，それぞれに接続可能な助動詞などを列挙してある助動詞相互の活用形別接続テーブルと，助詞および助詞連続について接続可能な名詞，用言，助動詞を列挙した助詞テーブルからなる．
	カルテ文章処理のための文型チェック辞書	九州工大	上松　弘明	・循環器疾患のカルテの文章解析をする際に用いる．カルテの文章から臨床診断上重要な語句を出す時，9つの文型を設定し，意味的に正しい語の組

	データ名	組織		備考
				み合せが列挙されている.
	拡張された文節を構成する要素的表現辞書	福岡大	首藤 公昭	・日本語文を構成する単位を意味機能のうえから網羅的に抽出・整理した辞書．ただし，接頭語，自立語の大部分は除く．多数の慣用的表現が含まれており，付属語的表現・接尾語的表現・自立語的表現に分類される表現が収録されている.
	属性マスタファイル	NEC	鵜木	・日本語処理システムで使用する文字に対して，部首，画数，音読み，訓読み，および新字，旧字などを区別する文字区分などの属性情報が，1文字単位の辞書データとして蓄積されている.
	姓マスタファイルおよび名マスタファイル	〃	〃	・姓名のカナ漢字変換を行うために用意された，姓名の読みがなに対応する漢字姓名データである．姓名の使用頻度も把握されている.
	漢字企業名マスタファイル	〃	〃	・企業名のカナ漢字変換を行うために用意されたもので，企業名を構成する単語別に蓄積したファイルである.
	付属語表	防衛庁	島貫 隆光	・用語の語尾，助詞，助動詞，形式名詞，接尾語，補助動詞等が収録されており，自立語，付属語との接続関係で75分類されている．名詞の格，動詞の態の区別，文節末の形態などの情報が付与されている.
2	J-ANALYSER（仮名）の為の補助入力データ	電総研	池田 尚志	・J-ANALYSER（仮名）の為の付属語辞書，各種のルール等である．約110語の付属語に対して，活用，接続，整形，係り受け，係り結び，変形規則等の属性が与えられ，活用型は約50型，約15型の変形規則が用意してある.
	遺跡関係用語	奈文研	田中 琢	（前述 1類の備考を参照）.
	教育文献から抽出されたキーワード	京教大	永野 和男	・教育文献表題から抽出された名詞リスト，および，それらの機械的操作による構造．対象文献は1,693で，抽出名詞は2,897である.
	企業名単語ファイル	日本ユニバック	田中 康仁	・企業名の単語がカナと漢字でファイルされており，データの件数は約12万件である.
	姓ファイル	〃	〃	・日本人の姓を集めたファイルで，データの件数は12万7千件である.
	名ファイル	〃	〃	・日本人の名を集めたファイルで，データの件数は26万件である.
	新コンサイス英和辞典	特定研究総括班「言語」		・新コンサイス英和辞典の全文を，ほぼそのままの形式で入力．現在校正作業が進行中．訳文，説明文等にあらわれる日本語文はわかち書きされている.
		京 大	長尾 真	
	表題文翻訳のための辞書	京 大	辻井 潤一	・表題文の機械翻訳のために，JICST 文献速報磁気テープ中に含まれた約1万の英語に対して日本語訳があてられている.
	漢字の字種	IBS	木村 睦子	・国立国会図書館蔵書目録，書名索引（S. 23-33）のうち，書名のみを対象に，当用漢字および人名用漢字以外の漢字の字種および頻度を調べ，頻度順に表にしたもの．異なり字数1,937字，延べ字数約1万6千字である.
	同音異語リストと文脈データ	〃	〃	・国研報告「雑誌九十種の用語用字」の五十音順リストに出現する同音異語のうち，名詞どうしの組合せを拾い出した．全部で278組634語である．また，同音異語リスト記載の語について国研所有のカードを抜き出しコピーした．かなりの量の文脈がついている.
3	ひらがな連系頻度表	〃	〃	・漢字かなまじり文約10万センテンスから，ひらがな連系のみを抽出し頻度を調べたもの．延べ度数748,980，見出し連系38,639のうち，頻度2以下のものを除いて7,011見出しが掲げてある.
	雑誌九十種語彙・表記データ	国語研	中野 洋	・昭和31年に刊行された雑誌のうち九十種を対象として，国立国語研究所が行った語彙調査の結果データ．語彙と度数とその表記形のバラエティを含んでいる.
	国研漢字調査	日本 IBM	藤崎哲之助	・国語研究所の二つの漢字調査（現代雑誌，新聞）の頻度を機械可読したもの．1）雑誌頻度順，2）新聞頻度順からなる.
	日本文の1字および2字組の出現頻度	東 大	小野 芳彦	・国語研究所で採取された約400万字の新聞から日本語文の文字の1字および2字組の出現頻度を求めたもの.
	カナ文字列	日本ユニバック	田中 康仁	・JICST 文献の中よりひらがな部分を抽出し，この中でよく使われる文字列を集めたものである．約3万3千件の文字列ファイルである.
	日本語の構文規則と一次および二次の生起確率	IBS	木村 睦子	・構文実態調査に基づいて作成した約200個の構文規則とその生起頻度.
	文型表	防衛庁	島貫 隆光	・単文（述語一つの文）の構造は，述語が定まればそれに直接かかる名詞の格や意味区分が定まる．それを類型化したものが文型パターンで，約5,600語を586に分類，分型は1,686に分類されている.
4	Syntactic Rules Data (SRD)	大府大	西田富士夫	・事実関係を記述する基本的な構造をもつ文，文献標題や特許請求範囲文などの名詞句や名詞節を対象とする文法規則データで，文脈自由規則の集まりと品詞間の順位関係表からなる.
5	現代新聞語彙調査の KWIC 索引	電総研	植村 俊亮	・国立国語研究所が1965年いらい行ってきた現代新聞語彙調査データのうち，公開された約750,000長単位語にのぼるデータの完全な KWIC 索引である.

	データ名	組織	備考
5	中国語白話文の KWIC 索引	電総研　植村 俊亮	・中国白話文約100万字について漢字4けたの数字符号で表現されているデータに関して，漢字1字や2連字の出現頻度調査を行っている．また，文法情報投入作業も行っている．
	教育文献表題	京教大　永野 和男	・教育学研究・全論文，教育心理学研究・戦後全論文，教育心理学全国大会発表論文など，教育分野に関する文献表題，著者名，文献名，発表年が収録されている．
	教育目標，学習項目	〃　〃	・教育現場で用いられている教育目標，学習項目．(1) 小学校1～6年の算・社・理の学習項目，2) 中学校数学1～2年の教育目標の全単元，3) 高校物理Iの学習項目）が収録されている．
	法律文データ	電通大　岡本 哲也	・大気汚染防止法，公害紛争処理法，大気汚染防止法の一部を改正する法律，大気汚染防止法および水質汚濁防止法の一部改正などの法律文データが収録されている．
	自然言語分析データ	九　大　吉田　将	・自然言語の機械処理に関する研究を遂行する上で，随時必要となった資料を集めたもの．約140項目のデータ（A4紙 約17,000枚）を分類・整理中である．
	音声データ	北　大　伊福部 達	・日本語68単音節の50名による音声，札幌37学校生徒約10名の音声，北大・歯学部病院通院中の不正咬合者約10名の音声，九官鳥3羽の音声．

ある程度各研究グループの目的にあわせて作成されているために，完全に汎用的にはなっていないが，少なくとも見出し項目の選定・追加など，グループ間でデータ交換する利益は大きいと思われる．企業名のデータ，姓名のデータ等は，実用的なシステムを開発してゆく上で，貴重な資料である．

3　日本語統計データ：漢字の頻度調査，および「ひらがな連続」の調査等は，形態素処理のアルゴリズムおよびその時に使われる2のカテゴリのデータを準備する際の重要な資料となる．

4　文法：文法は本来 parser の構成に依存するので，汎用的にはなりにくいが，防衛庁で作成された「文型表」は，日本語の各動詞（形容詞）が，日本語文中でどのような格助詞を伴って表われるかを網羅的に収集したもので，文法としてよりも，各研究グループが parser や文法を作成する場合の基礎資料として貴重である．

5　その他：検索用に蓄積されている言語データ（京教大），機械処理のために日本語資料を人手で整理したもの（九大），KWIC の出力結果等を含む．

4. むすび

前にものべたように，今回の調査は現在日本でおこなわれている活動の一部をみることによって，現況を垣間見るようなものとなった．よせられた回答自体は，もっと詳細にわたるものであった．今後，別途印刷し出版できればと思っている．なお，今回の調査に関連してのご注意，あるいは調査もれのプログラム・データ等がありましたら，是非筆者等に連絡頂きたい．調査に協力頂いた各位に感謝します．

アンケート調査の資料整理に，多大の協力を頂いた久米雅子嬢に感謝します．

（昭和54年7月9日受付）