解　説

日本語の形態素分析†

中野　洋†† 野村雅昭††

1. はじめに

日本語には分かち書きの習慣がない．しかし，機械処理のみならず言語研究の分野でも文や文章を単語に区切る作業は必須のものである．ところで，何を一単語と認めるか，どの語とどの語を同語とし異語とするかは非常に難しい問題である．国立国語研究所では創立以来大量の語彙調査を行っているが，単語に区切る作業（単位切り作業と呼んでいる）と同語異語の判別作業はその最も重要で困難な作業である．

単位切り作業は一見自動化されやすいものと見られる．それは，我々の素朴な言語意識で語というものを漠然と認めているからであろう．しかし，自動化はもちろん，言語の上での定義さえも容易ではない．

たとえば，いまかりに「語とは，一つの事物の名，あるいは単位的な概念を表わすものである」と定義しよう．するとただちに次のような疑問が湧いてくる．すなわち，「春風」と「春の風」，「電子計算機」と「コンピュータ」，「ロバ」と「うさぎ馬」と「うさぎのような長い耳を持った馬」とは，それぞれ同じ事物・同じ概念を表わすから同じ語であると言ってよいのか．それらが，語の構造が異なるから別語だというのなら，「ヒノキ」と「松の木」，「タケノコ」と「男の子」と「僕の子」は同じ構造だから同じ語だと言ってよいのか．

ところで，日本語には単語がどれぐらいあるのだろうか．国語研究所の新聞の語彙調査[1),2)]の結果は次の通りである．

	延べ	異なり
長単位 (1/3)	679,342	101,081
長単位 (3/3)	1,967,575	213,368
短単位 (1/3)	940,533	47,805

すなわち，昭和41年の新聞（朝日・毎日・読売）から1/60の割合でサンプリングした標本のうち，1/3の量に当たる延べ約68万語（長単位）を調べると異なり語数が約10万になる．標本を全体（3/3）にすると，延べ約200万語で異なり語数は21万余に増える．つまり，標本をふやせばふやすほど異なり語数はふえる．いつ収束するのかわからない．

言語処理を考えると，処理の対象がふえると言語処理用の単語辞書もどんどんふやさざるをえなくなる．では，辞書のカバーする範囲をかえないで収録語数を小さくすることはできないか．それは可能である．これまで述べた語数は長単位[1)]と呼ばれるものについてである．たとえば，「日本語／の／形態素分析」と'／'で切られるような単位だが，これを「日本／語／の／形態／素／分析」のように切る単位（これを短単位[1)]と呼んでいる）で辞書を作れば，少ない語数で大きな延べ語数をカバーできる．先にあげた数字で言えば，標本の1/3のデータを長単位で数えると異なり約10万語が，短単位で数えると異なり約4万7千語ですむ，単位を短くすることにより，半分以下の数で同じ標本をカバーすることができるのである．

では，どこまで語の単位を短くすることができるか．

2. 言語単位

2.1 文法記述用の単位

それ自体意味を担っていて，かつ最小の単位は何か．それに対する言語学側の答は「形態素」ということになろう．

ブルームフィールドによると形態素は次のように定義される[3)]．

「他の何等かの『言語形式』と部分的に音声＝意味的類似のある『言語形式』を『合成形式』(complex form)とし，他のいかなる『言語形式』とも，部分的な音声＝意味的類似のない『言語形式』は『単純形式』(simple form)または『形態素』とする．」

「言語形式のうちで単独では発せられないものを『付

† An Analysis of Japanese Morpheme by Hiroshi NAKANO and Masaaki NOMURA (Depertment of Comutational Linguistics, National Language Research Institute).

†† 国立国語研究所

属形式』(bound form), その他のすべてを『自由形式』(free form) とする. 『最小の自由形式』(a minimum free form) が『単語』である.」

筆者らは, 日本語の形態素を次の三種に分けて考える. 言語形式のうち, 文節の構成要素とはなるがそれ自体実質的な意味をもたず構文上の機能にかかわる形態素を助辞とよぶ. 単語の意味的な中核となるもので, 単独で単語を構成することもできる形態素を語基と呼ぶ. また, 単語の構成要素として語基と結合して, 形式的な意味をそえたり, 語の品詞性 (文法的性格) を決定したりするが, 単独では語を構成することはできない形態素を接辞と呼ぶ[16].

学校で習う文法は橋本進吉の文法によっている. 橋本文法[4] では文を実際の言語としてできるだけ小さく区切った一単位を「文節」とよぶ. 文節は音節の順序・アクセントが一定し, その間に音の断止がない. 文節を構成する単位が語である. 語は意味・音節・アクセントは一定するが, 文節を作る時には変化することがある.

橋本の他, 山田孝雄, 松下大三郎, 服部四郎, 時枝誠記なども文法記述のための言語単位[5] を提出している.

2.2 語彙調査用の単位

語彙調査においては, 単位の長さおよび質が一定していなければならない. たとえば,「進歩的」が一単位で「建設的」や「安全性」が二単位では困る. 辞書によって単位切りすれば単純明解だが, 完全な辞書は現在ないし将来にも現われないと思われる. また, 辞書が人間にとって引きやすい形になっている以上, これによって切るのは問題がある.「花より団子」のような連語,「蚊取り線香」のような複合語,「山」のような単純語, 最近は漢字見出しまであるのが辞書である.

国語研究所では, 誰が単位切りしてもゆれのない語彙調査用の単位として, 文節単位の考えを基にした α 単位, 現代語において意味を担う最小単位を基にした β 単位を設定した.

次に α 単位・β 単位の概略について述べる.

α 単位[6] は, おおむね品詞論を前提とした分割をおこない, それに該当しないものについて主として意味の関係から分割した.

まず, 品詞論を前提とした分割では, ①助詞・助動詞あるいはその連続の後, ②用言の中止法・終止法・命令法の後, ③用言の連体修飾法・形容詞の連用修飾法の後, ④連体詞・副詞の後, ⑤接続詞の前後, ⑥感動詞の後, ⑦体言の独立格の後で分割する.

意味の関係からは次の場合に分割する. ①並立する語, ②体言が, 属性概念を示す語の前にあって, その属性概念を賓とする主概念となっている場合 (例, 事態／急迫) および, その属性概念を補充する関係にある場合 (例, 操業／開始), ③「～用」「～向」「～式」などで導かれる連体修飾格に立つものとこれによって修飾されるもの (例, 男児用／外出着), ④全体と部分の関係に立つもの (例, 東京／大井海岸), ⑤行政区画とか各種の機構などの各段階を表わす部分, ⑥数量の単位の変わるところ, ⑦数量を表わす語の前, ⑧時間・分量・程度・方向・位置・関係などを表わすもので接尾語的に用いられたもの (例, 入営／直前). ただし, 体言にじかに接続したサ変動詞 (例, 活躍する) などは切り離さない.

その他, 動植物名, 十干・十二支・固有名詞などは切り離さない.

β 単位[7] は, 最小単位が, ある条件を満たす形で結合した (または結合しない—0回結合とみなす) 結合体である. ここで, 最小単位とは, 現代語として意味を担っている最小の言語単位をいう.

和語の場合, それ以上分割すると意味がわからなくなるか, もとの意味がなくなるかする, その分割それぞれを一最小単位とする. 漢語の場合, 原則として漢字一字で表わされる部分を一最小単位とする. 漢語以外の外来語の場合, 原語一語を一最小単位とする. β 単位を規定する必要上, 最小単位を以下のように分類する.

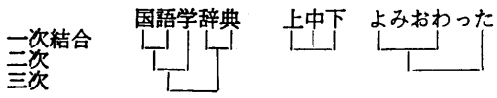
- 一般　山(やま) 人(ひと) 社 会 動く また
- 数　一(いち) 千(せん) 一(ひと) 二(ふた); 幾(人) 何(回)
- 付属要素　お ども 的 がたい (行き)かける
- 符号　イ 甲 A ○○(まるまる) = ×
- 助詞・助動詞　です た も から など

助詞・助動詞の範囲は, 大体普通の文法書にいうものである. 形容動詞の語尾は独立の助動詞 (または助詞) と認める.

付属要素は, 接頭語・接尾語・形式名詞・補助用言などからなる. 一般的な定義を下すのはむずかしいので, すべてを表にあげる.

最小単位の結合とは, 文節内での言語要素の結びつきのことである. 更に, その結びついた一まとまりが

意味・機能の上でも一まとまりになるという条件がつく．結合の二三の例を示す．



β単位作業規則

(1—1)　人名は姓を1β，名を1βとする．

(1—2)　行政区画を表わす地名は「都，府，県，郡……丁目，番地」を除いた部分を1βとする．

(1—3)　国や地形などの名前で，類概念を表わす部分は，地名に含めない．

(2—1)　数は他の最小単位との間を区切る．

(2—2)　数の間どうしの結合は，一，十，百，千の称えをとる桁ごとに1βとする．10^4の累乗を表わす部分（万，億，兆など）はそれだけで1βとする．

(3)　付属要素，符号，助詞・助動詞は，一最小単位を1βとする．

(4—1)　個々に独立して使われない最小単位の並列，ローマ字を並べた略称，区切るべき位置に問題のあるもの（|○○|○|か|○|○○|かが決し難い場合），第二次結合まで有する略語のうちその二次結合をしている前部分を置き換えて他の類似（下部）団体をさすことがないもの，別に示すものはそれで1βとする．

(4—2)　建物・会社・店・学校・乗物・団体・宗派・流派・新聞・雑誌・商品・人種・民族・言語の名前で，種差を表わす部分が一最小単位である時，その部分を1βとする．

二最小単位の一次結合体の長さが，外来語どうしでは七音節，その他の結合では六音節を越える場合の，それぞれの最小単位を1βとする．

個々に独立しうる最小単位の三つ以上の並列は一最小単位を1βとする．

(4—3)　先にあげた最小単位の分類のうち，「一般」の分類に入れた最小単位二個の一次結合を1βとする．

(4—4)　以上の規則によって認められたβ単位に，前または，後ろから順次に結合した一最小単位は，それだけで1βとする．

(4—5)　他と結合しない一最小単位は1βとする．

3. 分かち書きプログラム

日本語の普通の文章は漢字かな交り文で書かれ，分かち書き* されない．したがって日本語を機械で処理する場合，最初に「分かち書き」処理が行われなければならない．分かち書きプログラムはこれまで数多く作られた．その方法はテーブル方式とプログラム方式とに分けられる．テーブル方式は大きな辞書を用意しこれによって分割しようとするもの，プログラム方式は原則として辞書を使用せず，簡単な文法情報や語結合の法則などを手がかりにするものである．前者は辞書の作成に難点があるが，システム改良の必要が生じた時，辞書とプログラムとが分離されている方が独立な改良ができる，種々の方式による実験ができるという長所がある．後者は実用化に近く，処理速度も早いが，一般に適用範囲が限られており精度もやや低い．

前者に近い方法として，石綿敏雄らの単語認定プログラム（AUTOSEG）[8]がある．これは単に自動分割だけでなく，構文解析処理も行っている．また，文法規則の妥当性を検査するため可能な解をすべて出力する．それだけ，実用から遠ざかっている．後者に近い方法として，江川清[9),10)]，坂本義行[11)]等のプログラム（後述）がある．しかし，まず後者の方式で切り，前者の方式で修正するという両者の融合方式[12)]や構文解析の中に分かち書きを融合させるという方式[13)]が有効であろう．

目標である単語の長さも，文節のような長い単位と形態素のような短い単位がある．文節単位は直接文を構成する単位であるから構文解析の処理単位となりえる．しかし，テーブル方式でこれを処理する場合には，テーブルが大きくなる．形態素単位はテーブルは小さくてすむが，文節単位への合成が必要になる．長い単位の分析は意味処理にかかわる部分が多く，構文解析と似た困難が生じる．意味処理がどのみち必要になるなら，この段階で始めるのも一つの方法であると思われる．

3.1 長い単位切りの問題点

漢字かな交り文は分かち書きされていないが，字種の違いが視覚的に単語の切れ目を示す場合が多い．送りがなの処理ができれば，字種のかわり目を単語の切れ目とするとほぼ長い単位が得られる．ひらがなから非ひらがな文字へ変わるところを切れ目とするとほぼ文節単位が得られる．

坂本[11)]は特許公報27万5千字について字種の変わり目（ひらがなから他の字種）で分割した場合には，どの程度の誤りが発生するかを調べている（漢字連続

* 文を単語に分割することを「分かち書き」「言語単位分割」「単語認定」「単位切り」などという．

の中は調べない)．それによると，同一字種内で切れるべきなのに切れていない(10.7%)，字種の変わり目で切れるべきなのに切れていない (3.5%)，切れるべきでないところで切れている (5.8%) となっている．これらを修正処理するために少量のテーブルと文法情報・語接続情報を利用している．

字種の変わり目が単位の切れ目でない場合，同じ字種の連続が単位の切れ目である場合は次の場合が考えられる．すなわち，前者は，送りがながつく語（例，動く，美しい，最も）複合動詞およびその転成名詞（例，置き換える，引き出し），まぜ書き表記（例，さく酸，まま母，ご迷惑，朴とつ，肺ガン，パイ煙，5カイリ，2ケ国，1か月，サボる）などであり，後者は，形式名詞・補助用言・副詞・連体詞・接続詞などひらがな書きされやすい語が他のひらがな書きの語に接続した場合（例，さてこのことをもっと押し進めていくと），漢字書きされやすい副詞・時や数量をあらわす名詞，形式名詞，接続詞，連体詞が漢字書きされた語に接続する場合（例，或人，元首相，昨日東京に到着した，五円値上り，その後先生は，結局提出した，又始まる）などである．

3.2 短い単位切りの問題点

長い単位にくらべ短い単位は自動化がむずかしい．問題は複合語，特に漢語の長い漢字連続をどう切るかである．たとえば，「磁気記憶装置」は「磁気」「記憶」「装置」に分かれる．このことから漢字二字を一単位とする方式が考えられるが，「未完成交響曲」の場合は「未」「完成」「交響」「曲」と分かれ，二字ずつ切ったのでは正しく処理できない．そこで，「未」「曲」のような接辞（あるいは，統計的に得られた接辞として使われやすい漢字）のリストを利用し，これを切り出した後に二字で切るという方法[10]が考えられる．しかし，長尾らの報告[12]によるとリストの利用は必ずしも精度を上げない．同様のことを坂本も報告[11]している（形式名詞のリストが悪影響をおよぼす)．たとえば，「元」「時」のリストにより，「元／首相」「同位／元／素」「この／時／彼は」「この／時／計は」などのように，正しく分割される場合と誤る場合とがおこるのである．以上の方法はプログラム方式である．

テーブル方式による短い単位の切り出し[12]も行われている．最長一致法による分割である．辞書による分割の問題点は次のとおりである．

i）辞書に語が登録されていない場合……先に述べたように完全な辞書は用意されえないだろうから，そのための処理は考えておかなければならない．専門用語や新造語をいかに有効に登録するかも問題になる．後に述べる語形変化に対する処理も必要だ．特に，活用語の登録語形への変換は必須である．省略形の完全形への変換「電磁界→電界磁界」は自動化が難しい．省略形も辞書に登録する必要があろう．

ii）辞書に登録されている語で二通りの分割が可能である場合……「不安定」は「不安」と「安定」の登録語によって解が二通り出る．一字で表わされる語が登録されている場合，解が多く出る可能性がある．

3.3 言語処理用の言語単位

α 単位，β 単位は語彙調査用の単位であり，文節単位は文法記述用の単位である．言語処理を考える場合，必ずしもこれらの単位に従う必要はない．目的に応じた効率のよい単位であるべきだ．

坂本[11]は文節に対し，次のような規準を加えている．すなわち，形式名詞は付属語とする，普通名詞＋数詞は分割しない，形式用言は狭い範囲に限定する，特殊記号は種類によって字種を決めるとしている．これにより，「高クロム合金に対し／約 0.2% 以下の／ケイ素含量を／保つことが／困難であるような」のようなものを文節としている．

長尾ら[14]は，「文法上の文節と一致しないが，単語レベルの接続関係が境界を越えて影響を及ぼすことはないと考えられる」から，ひらがなから漢字に変わるところを文節の切れ目としている．また，これより優先して慣用句の切り出しを行っている．これは，機械翻訳という目的のためには，慣用句を分割して品詞の認定を進めるのは，いたずらに解釈を複雑にするだけであるからである．慣用句は首藤公昭[15]があげた科学技術論文にあらわれる付属語表現を新たに整理したものである．Aspect, Tense の形式素（～ている，～てしまう)，Modal 形式素（～はずである，～かもしれない)，文間の関係を表現する形式素（～からである，～わけである)，Case の形式素（～によって，～という目的で）などがそれである．

中野は国語研究所における文脈つき用語索引の作成，語彙調査等のための自動処理プログラム[24]を試作した．これは自動単位切り，読みがなつけ，品詞認定等の機能を持っているが，その結果はまだまだ言語研究に堪えるものではない．実用化を考えるなら，人間と機械との随時相互の情報のやりとりを可能にする機能が必要である．これは言語学側の機械利用の一つの

アプローチである．

筆者らは，単位切りの問題を文字や文字列のレベルで処理しようとする限り完全な自動化は難しいと考える．結局は文の構造，文章の意味まで問題にしなければならないだろう．1章であげた例や次の例は，それらが解決されなければ正しい処理はできないだろう．

都区内（都内・区内の省略形である）

諸国民（諸々の国の民か，諸々の国民か）

米ソの緊張は極度に達した（この場合は「極度／に」と切る．「極度に／おびえる」と区別する．）

運転できた（「運転出来た」と「（自分の）運転で来た」の場合を区別する．）

大の字になる（この場合は一語．「国語の時間に大の字を習う」の「大の字」は三語に分割する．）

二枚目を用意する（「俳優は」と「皿は」によって切り方が異なる．）

法案を審議中の国会（「法案」は「審議（する）」にかかり，これに「中」が結合した形．ただ単位切りするだけでは問題の解決にはならない．「試験を採点の先生方」「会議をお休みの場合」なども同じ．）

4. 語構成

自動単位分割において困難な問題は長い単位（複合語）を短い単位に分割することである．その一般的な方法を示すことは残念ながら出来ない．ここでは語がいかに構成されているか[16]について述べる．

日本語の形態素はそれが単語の構成要素となるか否か，単語の意味的な中核になるか否かによって，語基・接辞・助辞に分けられる．しかし，語基と接辞，接辞と助辞・活用語尾などの境界はそれほど明確ではない．ことに漢字一字で表される字音単位には，問題となるものが多い．漢字一字に相当する形態素は，意味的には語基相当の単位であるが，ほとんどが自立できず，他の字音形態素と結合して安定した単位となる．つまり，漢字二字が結合して一語基相当の機能を持つので，これを一語基とするのが実際的である．二字熟語でも，「国際・民主・合理」など自立しない語基もある．

「乗用車」の「車」を接辞とするなら，「汽車」の「車」も接辞とすべきではないか．「酸性」と「国際性」の「性」はいずれが語基でいずれが接辞かなど，結局，一律に語基あるいは接辞とするのは無理があり，意味が実質的で明確なものを語基とし，形式化したものを接辞とするのが穏当な措置といえよう．

語は次のように分類される．

- 語（単語）
  - 単純語（語基一個）
  - 合成語
    - 派生語（語基＋接辞）
    - 複合語（語基＋語基）

語種（それがもともとどの言語に由来したかの別）によって語の構造に差異が認められることがある．以下では，語種の差を示すために，和語はひらがなで，漢語は漢字で，外来語はカタカナで表わすことにする．

4.1 単純語

「やま」「あそぶ」「肉」「建設」「スポーツ」「スマート」などが単純語である．

単純語は語基一つによって構成されている．語基がどのようにして生まれたかはここでは問題にしない．

ただし，漢語については漢字二字の結合したものを一語基とする．その構造を分析できるものもあるが，合成語の構造とはかなり異なること，一語意識が強く分割がむずかしいものがあること，省略形が多いこと，無意味形態素が存在することなどの理由によりここでは問題としない．

4.2 派生語

派生語は，接辞と語基の結合形であるから，一次結合にかぎれば，つぎの二つの形式しかない．

①接頭辞＋語基　お一なか　こ一にくらしい　無一関係　超一特急　ポスト一三木

②語基＋接尾辞　うつくし一さ　かあ一さん　健康一的　必要一性　スラム一化

①のタイプは種類・量ともに多くない．特に和語系の接頭辞は，「お（御）」「おお（大）」「はつ（初）」などいくつかを除けば派生語の生産性は低い．漢語系の接頭辞は種類も多く生産性も高い．新聞の調査に出現した接頭辞的な一字漢語は250種類ある．これらは，すべてが純接辞的なものとは限らないが，その用法は現代語で発展したものでその多様さは和語を圧倒している．

その中の「無（一条件）」「不（一利益）」「非（一現実的）」「未（一発表）」など否定の意味を添える一群の接頭辞は，二字漢語の要素としても使われるが，和語とも結合し，接辞的な性格をそなえている．これらはまた，結合形全体にいわゆる形容動詞の語幹相当の品詞性を与える．

漢語系の接頭辞で，最近よく用いられる「反（一体制）」「超（一高層）」「過（一保護）」などは，一字の

意味としては，動詞性の強いものである．

②のタイプの接尾辞には，単に意味を添えるものと，結合対象となる語基の文法的性質をかえるものがある．前者は「さん（課長ー）」「こ（カギっー）」など待遇関係や人間をあらわすもの，「員（研究ー）」「業（運送ー）」など語基と接辞の中間的な性格をあらわすもの，助数詞的なものなど種類が多い．

②の文法的性格をかえるものには

〔名詞をつくるもの〕 さ（暑ー），け（寒ー），み（強ー），性（可能ー），イズム（ゆっくりズム・がんばりズム）

〔形容動詞の語幹およびそれに準じた体言をつくるもの〕 げ（うれしー），そう（寒ー），やか（つやー），的（法ー），性（酸ー），風（王朝ー）

〔形容詞をつくるもの〕 い（赤ー），しい（とげとげー），ぽい（あきっー），らしい（男ー）

〔動詞をつくるもの〕 る（デモー），する・ずる・じる（かろんー），めく（いろー）

〔サ変動詞の語幹をつくるもの〕 化（液ー，近代ー，ドラマー），視（敵ー，問題ー）

〔副詞をつくるもの〕 に（特ー），上（事実ー）

などで，これ以外にもかなり多い．

4.3 複合語

語基の結合関係を文の成分相互の関係に対応づけて考えるために，語基を次の四種類に分ける．

〈A類（体言類）〉 やま，人間，文化，テレビ

〈B類（相言類）〉 あお（い），うれし（い），別（な），貴重，スマート

〈C類（用言類）〉 ね（る），やすみ(む)，発見，スタート，ミックス

〈D類（副言類）〉 また，一斉，突然

次に複合語のパターンを示す．

〔複合動詞の構成パターン〕

① A＋C……Aガ（＝ヲ・ニ・デ）Cスル

目ーざめる，名ーづける，夢ー見る，耳ーなれる

② B＋C……Bノ状態デ（＝ニ）Cスル

青ーざめる，遠ーのく，長ーびく，若ーがえる

③ C＋C……Cシタ状態デCスル

撃ち一落とす，掃き一出す，踏み一抜く

複合動詞はいずれも動詞が後にくる．③が最も多い．

〔複合形容詞の構成パターン〕

① A＋B……Aガ（＝ニ）Bノ状態デアル

幅ー広い，名ー高い，興味ー深い，耳ー新しい

② B＋B……Bノ状態デBデアル

暑ー苦しい，甘ー酸っぱい，細ー長い

③ C＋B……Cシタ状態デBデアル

蒸しー暑い，まわりーくどい，こげーくさい

④ D＋B……Dノ状態デBデアル

ひょろー長い，うすらー寒い，むずーがゆい

複合形容詞の後部分は，〈語基〉＋〈接辞〉という構造をもつから，厳密にはこれは合成語である．

〔複合名詞の構成パターン〕

(第一類)

① A＋B……AガBノ状態デアル

色ー白，身ー軽，胴ー長，物価ー高，栄養ー豊富

② C＋B……CスルコトガBノ状態デアル

話ーべた，待ちー遠，望みー薄，実現ー可能

③ A＋C……Aガ（＝ヲ・ニ・デ・ト……）Cスル

雨ー上がり，動脈ー硬化，川ー遊び，ドイツー製

(第二類)

④ B＋C……Bノ状態デCスル

早ー起き，急ー上昇，完全ー消毒

⑤ C＋C……Cシタ状態デCスル

立ちー読み，見ー習，徐行ー運転

⑥ D＋C……Dノ状態デCスル

ほろー酔い，再ー検討，一時ー停止

(第三類)

⑦ B＋A……Bノ状態デアルA

丸ー顔，若ー物，甘ー納豆，温暖ー前線

⑧ C＋A……Cスル（＝シタ・シテイル）A

打ちー傷，流れー作業，救援ー投手

(第四類)

⑨ A＋A　川ー道　核ー兵器　大学ー病院

(第五類)

⑩ A・A　朝ー晩　足ー腰　竜頭ー蛇尾

⑪ B・B　あまーから　すきーきらい　自由ー自在

⑫ C・C　売りー買い　読み・書き　比較ー対照

(第六類)

⑬ A＝A　ひとーびと　ところーどころ

⑭ B＝B　ながーなが　はやーばや

⑮ C＝C　とびーとび　ちりーぢり

上のパターンに属する語数の多いものをあげると，和語が⑨，③，⑤，⑧，⑦，①となり，漢語が⑨，③，⑤，⑦，⑧，①となる．これだけで全体の約 95% を占める．⑨だけで和語では 40%，漢語では 30% を占める．

語基の意味的な結合関係は文の成分の関係と類似している．

第一類は格関係，第二類は連用修飾関係，第三類は連体修飾関係に相当する．③⑧は狭義の格関係のほか，いろいろなタイプがある．

複合名詞の分析のためには，単に語基の統辞論的な関係をてがかりにするだけでなく，語基の語彙的な意味をも問題にしなければならない．たとえば，「支配一人」と「使用一者」は⑧のパターンに属し，見かけは同じであるが，前者は「支配スル人」であるのに対し，後者は「Xガ人$_1$を使用スル，ソノ人$_2$」($人_1=人_2$)のXと人$_1$を消去したものとみなければならない．同様に⑨は，前部分の語基をP，後部分の語基をQとすると，

Qガ<場所>Pデ（＝ニ）VスルソノQ

Q_1＝自然物　V＝存在

山一ゆり　海一がめ　高山一蝶

Q_3＝現象　V＝生起

川一風　山一火事　都市一公害

Q_2＝生産物　V＝使用　XガPデQヲ使用スルソノQ

山一刀　ビーチーパラソル　雪上一車

などと分析できる．

なお，ここでは触れなかった日本語の複合名詞の生成文法について述べた奥津敬一郎の論[17]は機械処理にとって有効であろう．

5. 語形変化

形態素が合成するとき，あるいは文法的な性格がかわるときに語の形をかえる．

5.1 形態音韻変化

形態素が合成して，合成語を作るときに語形がかわる．これを形態音韻変化という．日本語には次のような現象がある．

連濁　芝居＋スキ→芝居ズキ

音便　ヒキ＋ハガス→ヒッパガス

フミ＋ハル→フンバル

母音の交替　アメ＋カッパ→アマガッパ

待機音の挿入　オトコ＋フリ→オトコップリ

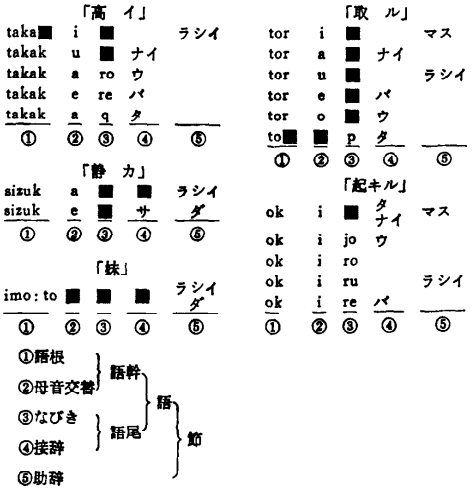
シリ＋アイ→シリヤイ

音節の脱落　カワ＋ハラ→カワラ

音韻の縮約同化　テ＋アライ→タライ

連声の現象　銀＋杏→ギンナン

これらの現象はある程度まで規則化できるが，例外が多いので完全に自動処理化するのは難しい．



5.2 活　用

活用とは語が文の中で使われる時，文法的な性格にしたがい語形をかえる現象である．語がどのような形で文中に使われるかについて宮地裕[18]の論を紹介しよう．

図中■はゼロ記号である．いわゆる音便形の子音の脱落現象である．①は変容しない部分，②は変容する母音部分だが一段活用「起キル」だけは変容しない．③は「ル・レ・ヨ・ロ」などの付加部分であって，江戸時代の国語学用語「なびき」（靡）を利用した．①②を語幹，③④を語尾とするが，②と③の所属は語幹・語尾のどちらにも入れられないわけでなく，また④⑤の間も分類のしかたによって異なる．

6. おわりに

分かち書きされていない日本語の文章からどのようにして言語単位を切り出すか．また，それらはどのような構成をしているかについて，これまでの研究を紹介した．語分割・語構成分析は人間にとっても困難な作業であり，その自動化に対し有効な情報を提供しえないのが現状である．しかし，自動化の目的を明確にし，それに合った単位を設定し，その適用範囲を限定する，あるいは人間の機械への情報付与を可能にするシステムを考えるなら，かなりの自動化も可能であると思われる．言語学と工学との協力がこれからの課題であろう．

紙幅の都合で，あるいは筆者の知見の不足によって触れなかった先学の業績も多い．お許し願いたい．

参 考 文 献

1) 国立国語研究所: 電子計算機による新聞の語彙調査, p. 342, 秀英出版, 東京 (1970).
2) 国立国語研究所: 電子計算機による新聞の語彙調査, p. 530, 秀英出版, 東京 (1973).
3) ブルームフィールド, 三宅 鴻・日野資純訳: 言語, p. 947, 大修館書店, 東京 (1962).
4) 橋本進吉: 国語学概論, p. 380, 岩波書店, 東京 (1946).
5) 服部四郎・川本茂雄・柴田 武編集: 日本の言語学第四巻文法II, p. 671, 大修館書店, 東京(1979).
6) 国立国語研究所: 現代語の語彙調査婦人雑誌の用語, p. 338, 秀英出版, 東京 (1953).
7) 国立国語研究所: 現代雑誌九十種の用語用字, p. 321, 秀英出版, 東京 (1962).
8) 石綿敏雄・斎藤秀紀・木村 繁: 言語単位分割自動化の研究, 計量国語学 No. 50, pp. 24-38 (1969).
9) 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 No. 44/45, pp. 46-52(1968).
10) 江川 清: 単位分割自動化のシステムについて, 計量国語学 No. 51, pp. 17-22 (1969).
11) 坂本義行: 文節単位の自動分割法—字種と平仮名連系による—, 計量国語学 Vol. 11, No. 6, pp. 265-276 (1978).
12) 長尾 真・辻井潤一・山上 明・建部周二: 国語辞書の記憶と日本語文の自動分割, 情報処理 Vol. 19, No. 6 (1978).
13) 田中穂積・佐藤泰介・元吉文男; 自然言語処理のためのプログラミングシステム—拡張 LINGOL について—, 電子通信学会論文誌, Vol. J 60-D No. 12, pp. 1061-1068 (1977).
14) 長尾 真: 計算機による日本語文章の解析に関する研究, p. 212, 文部省科研費報告書 (1979).
15) 首藤公昭: 日本語における文節の構造とその処理について, 福岡大学研究報告第35号 (1978).
16) 野村雅昭: 複次結合語の構造, 国立国語研究所報告 49, pp. 72-93 (1973).
17) 野村雅昭: 否定の接頭語「無・不・未・非」の用法, 国立国語研究所論集 4, pp. 31-50(1973).
18) 野村雅昭: 三字漢語の構造, 国立国語研究所報告 51, pp. 37-62 (1974).
19) 野村雅昭: 四字漢語の構造, 国立国語研究所報告 54, pp. 36-80 (1975).
20) 野村雅昭: 造語法, 岩波講座日本語9, pp. 245-284 (1977).
21) 野村雅昭: 接辞性字音語基の性格, 国立国語研究所報告 61, pp. 102-138 (1978).
22) 奥津敬一郎: 複合名詞の生成文法, 国語学 101, pp. 33-48 (1975).
23) 宮地 裕: 日本語の文法単位体, 岩波講座日本語 6, pp. 1-31 (1976).
24) 中野 洋: 言語処理のための一貫処理の研究, 国立国語研究所報告 61, pp. 17-40 (1978).

(昭和54年8月14日受付)