計算言語学 23－4
（1980. 6. 13）

漢字列長単位用語の抽出

田中康仁　日本ユニバック㈱

はじめに

日本語の分かち書きの問題は一見単純にみえるがいろいろな困難な問題を含んでいる。"分かち書き"を行うには，どのような方針で，何を目的にするかによって方法が大きく異ってくる。今までの用語の研究は国研の新聞，雑誌，小説が主なものであった。また国研の研究はどちらかというと短単位に分割するものである。これは，辞書の編集，語の意味分析，語基の研究等については，短単位の用語がよいが，機械翻訳のための用語，専門用語（熟語）を考えると長単位に分割した用語のほうが取り扱い易い。

日本科学技術情報センター（ＪＩＣＳＴ）でも用語の研究は行われているが，文献の検索という目的から少ない質問用語から，一致する情報を引き出すため，用語は短かく切られて処理することのほうが多い。

このような状況であるため，専門用語の収集は，学術用語辞典とかＪＩＳの用語，各種協会，学会，企業が出版している用語集程度のものである。これらは，すべてのものを集めたものではなく，特徴的用語の収集である。

将来，情報交換が急速に進む社会になるならば，機械翻訳，機械支援による翻訳が重要な意味を持ってくる。良い訳語や，用語の多義性を減らすには，長い用語の収集，研究が必要である。

ここでは，ＪＩＣＳＴのファイルを利用し網羅的に漢字列を収集した時に発生する問題点，用語の収集を自動的に処理する方法等について述べる。

1　長単位の用語について

これまでの用語の研究は辞書の編集とか情報検索のために行われてきた。このためには短単位を採用するほうがよかった。この理由は次のような理由からである。

1）短単位に用語を切ると，語の持っている基本的意味が抽出できる。

2）用語の数が多くならない。辞書等への編集が簡単になる。

3）情報検索では短単位に用語を切るほうが，質問者からの検索向上につながる。

このため多くの研究は用語を短単位に分かち書きすることが行われている。

このような考え方には次のように反論することができる。

1）短単位に用語を切ることによって，用語の意味を分析し利用することができる。

しかしこれは一部分の狭い範囲に適用できるわけであって少し広い世界ではこの考えはあてはまらない。計算機にこの小さな範囲のルールを入れて少し大きな世界を作ろうとすると，ルールだらけになり混乱におちいる。

例1　直角三角形，正三角形，鋭角三角形，鈍角三角形などは，三角形をより詳しく説明したものである。
この直角，正，鋭角，鈍角などによって意味がつかめる。

しかし，この考え方を四角形に拡大すると必ずしも当てはまらない。

例2　台形，四角形，ひし形，平行四辺形，長方形，正方形
これらは表記形式が"形"一文字を除いて，すべて異り，これらの文字列からだけでは四角形であるかどうかさえ判別できない。まして意味を統一的に抽出することもできない。

機械翻訳の辞書を考えるにあたっては，長単位の用語辞書でなければならないことが，次のような例でもわかる。

例3　原動機付自転車←→モーター・バイク←→motor bike

速乾性筆記器具←→マジック・ボール・ポイント・ペン←→magic ball pointpen

潜水用水中呼吸器←→アクアラング←→aqualung

懸垂式多灯型照明器具←→シャンデリア←→chandlier

傾動荷枠後扉開閉装置付特殊自動車←→ダンプカー←→dump car

このように，用語の短単位への分かち書きが他の言語との対応を考えるにあたっては意味がないことがわかる。例3は少し偏よった例であると思われるが，通常の専門用語の中からも長単位の用語のほうが良いことがわかる。

例4　接触性伝染病　contagious disease（医学）

接触性皮膚炎　contact dermatitis（医学）

接触制御　touch control　（電気）

接触抵抗　contact resistance（電気）

この例4からもわかるように，同じ“接触性”，“接触”でも英語の同一語とは対応しない。日本語の概念と接続その表現は，他の言語の概念と接続その表現は，必ずしも一致しない。このほか長単位の用語は概念が限定されるため，あいまいさが少ないとか，機械翻訳の際によりよい訳語を引き出せるなどという利点がある。

2）長単位の用語を集めることは収集の労力，ファイルの増大化が問題になるが，ある分野を限定しJICSTの機械可読ファイルや，特許ファイル，出版印刷の写植用ファイル，新聞社の写植用ファイルから組織立って集めれば問題ない。また，最近では磁気ディスクの容量が急速に増大していることも考えれば，長単位の用語の収集，研究の環境が整ってきたといえる。

一つの分野で使用される用語の数はどの程度で満足できるかということは，どの程度の精度を要求するかによって決まることであるが，数十万件程度集めれば十分と思われる。これは今後の研究課題の一つである。

3）語基の研究，辞書の編集，情報検索等の分野では，短単位の用語の分析が重要な意味を持っているが，機械翻訳，機械支援による翻訳等の研究開発のためには，長単位の用語の研究が必要である。応用分野が異ってくれば研究方法も変えるべきであろう。

さらに機械処理の面から考えれば次の長所がある。

4）パーザーが楽になる。

長単位用語を取り扱えば文を構成する要素が少なくなり，処理速度が向上するし，短単位に分割するために起るあいまいさが減る。これは言語を機械で処理するために重要なことである。

2　漢字列の抽出実験

日本科学技術情報センターの抄録テープを利用し，抄録の中から漢字列を機械的に抽出し，分析した。実験内容は次のとおりである。

1）実験データ：日本科学技術情報センター抄録テープ（経営編）

2）処理日時：昭和54年4月 ～ 昭和54年9月

3）抽　出　方　法　：　文字種による機械的分かち書き

4）抽　出　デ　ー　タ　：　漢字部分のみ250,000件抽出

5）漢字列の種類　：　50,848件

6）作　成　資　料　：　頭文字順KWIC，末尾文字順KWIC（各1156ページ）

但し，同一漢字列は1件にまとめた。

25万件の漢字列をすべてリスト分析することは大変な労力がかかるし，同一漢字列を多くリストしてもあまり意味がないので，同一漢字列を1件だけ表示し，その漢字列が何件あったかを示した。

次にこの例を示す。

漢　字　列　（先頭文字順）

			頻度
需要者は欲せず住みよく環境を第一としている。	住宅団地	プロジェクトの販売宣伝はその点に留意	1
オ州の隣接2都市をモデルに商店，街路，車両，	住宅地	から成る回路網を作り流量，圧力，抵抗	3
効性とその公衆との関係という立場から，著者は	住宅地域	における図書館と協会，組織団体，企業	4
と世界各国空港で採用している実例，空港周辺の	住宅密集地	を避ける優先滑走路方式を説明。新しい	1
最近のわが国経済は，公共投資，	住宅建築	の高水準持続，個人消費の伸びに民間設	2
もかかわらずたいへんな好況でそれは公共投資と	住宅建設	の役割，在庫投資の異変，卸売物価上昇	3
小建設会社としてスタートし，現在米国第4位の	住宅建設会社	で，第1の建築資材メーカにまで成長し	1
べて1975年に37．4％，社会II文化政策と	住宅建設基金	は37．3％，生産発達基金は34％そ	1
増大等で6兆の資金調達，個人部門の資金余剰は	住宅投資	が盛行しても18兆の金融資産増で，株	1
の現状と今後の見通しについて述べた後，今後の	住宅生産	の工業化を一層推進していくための技術	1
・スーパーなどの個人消費関連，および不動産，	住宅産業	などである	4

漢　字　列　（末　尾　順）

			頻度
，国際特許情報検索委員会（ＩＣＩＲＥＰＡＴ），	国際学術連合会議	（ＩＣＳＵ），世界科学情報システム（	4
，宇宙技術の非宇宙業界での利用を促進するための	合同会議	の主催である。数回の準備的な会合の後	6
独ドイツ統一社会党第8回党大会および同党の第6	回中央委員会議	の決定を記し，徒弟の教育II訓練で文化	1
ータの収集，評価，提供に関するＣＯＤＡＴＡ第2	回会議	（1970年9月）の報告	3
説明し，例としてオトラドネンスク掘さく事業局の	科学的労働組織化集団会議	の活動状況について記述	2
，情報技術者ら1700人が参加したシンシナチの	国内情報会議	の議題と結論を中心とし，主としてＥＰ	1
年弱で軌道にのるまでの経過を紹介。ＱＣ勉強会，	職場会議	，発表会，懇談会などの勉強内容II対象	1
アーカイブが数多く設立された。国際的レベルでは	国際社会科学会議	や社会科学データアーカイブ会議などが	1
本年5月にボストンで開かれた，	産業衛生専門官会議	によって採用された，空気汚染濃度規準	1
議会報の1972年第612号報には，西独政府の	専門家会議	の注目すべき見解が掲載されていた。こ	2

このように同一漢字列をまとめると，次のような問題が発生する。

例1　研究する　→　サ変動詞の語幹

研究を続ける　→　名　詞

例2　早い　→　形容詞

早さ　→　名　詞

“研究”“早”という漢字列だけになると品詞情報がつかめないため意味がないという見方もあるが，用語の機械的抽出，その問題点を分析するためには，このような圧縮した資料が役に立つ。

漢字列の左右に文章を付けているため，特殊な漢字列が発生してもその原因を容易に分析することができる。また，膨大な漢字列を分析するためには，おおまかな分析から，より詳細なものへと移行すべきである。

3. 漢字列の分析

長単位の用語は文字種による分かち書きで十分用語の抽出が行われると考えがちであるが，次のような例がある。

漢　字　列

		頻度
を着用していれば衝突あるいは転倒した時にどの	位安全	1
の地域に多いか，新卒者は東京以外の地域にどの	位就職	1
本ミネチュアベアリング，3位カシオ計算機，4	位島野工業	2
工作所，島野工業，カシオ計算機が大幅上昇，1	位新日鉄	1
ラールを管理する唯一の方法は，その動的構造一	位相	4
先に筆者らが	位相型	1
ついての研究を報告した。すなわち，化学構造の	位相学的	1
なわち，化学構造の位相学的の記載，DARCの	位相学的分析	1
集合の間の関係を研究する。1つの例を上げて，	位置吸収集合	1
的適用法，実質サイズとMMC一最大実体条件，	位置度公差	1
このデータは正規化され凝視点の順序，期間及び	位置座標	4
になるかという観点からダミーを使って		1
するだろうか等につき調査報告する		1
でそれぞれの最近の輸出比率は51%，		2
，2位日産，3位トヨタ。利益ランクの		1
を究明し，確立することである。この場		4
データ構造の概念を提案し，意味空間が		1
の記載，DARCの位相学的分析，位相		1
，位相一情報の相関性，このDARC関		1
の構成は有限数のプログラム吸収集合の		1
)およびサイズの許容限界値の意義など		4
をMTに記録する。このデータにより凝		1

“位”で始まる漢字列の一部であるが，“どの位”，数字表現等のため単純な文字種の分かち書きでは用語の抽出は行えない。

次に，これら漢字列の前後に現われる問題点を場合分けして分析してみる。

3－1　漢字列の中で分かち書きが比較的やりやすい場合。

1）漢字で書かれた副詞が他の語と接続した場合

例　案外早く見つかった。　　若干悪くなる。

漢字で書かれる副詞は数が限られているため，これを表にすれば分かち書きに利用することができる。

2）文の中でしばしば使われる特殊な名詞が結びついた場合

例　上記物質，前記液体，当該専門家

“上記”“前記”などの特別な用語はあまり多くないため，これを表にし，分かち書きに利用することが出来る。

3）漢字列の最後に，“上，下，時，前，中，後，内，外”のような文字が接続する場合

例　基本上，反応条件下，分解時，溶接前，蒸着中，加工後，被覆内，発明成分外

これらは漢字の分かち書きを行うときに，これら文字を分離するか否か方針を立てなければならない。もし分離するならば，このような文字が用語の末尾にくる場合は多くないため，これを特別な表として処理することが可能である。“上下”“内外”などの用語を特別な表に入れ誤って分かち書きされないようにする。

4）次の漢字列が漢字列の末尾に来るときはその文字の例外表を調べ，該当しなければその文字を分離したほうがよい。

尚，或，且，毎，又，迄，蓋，該，為，及，第，等，目，間，約，例，各

例　解法及び結果を，図書館毎の，第四回目，商法第287条，振動等を，学生約120人，情報間の結合

このような文字が用語の末尾にくる場合は多くないため，3）と同様に例外表を用い処理することができる。

5）連体詞と漢字列

例　わが国領事が通報をうけず，その後発見の通告 ……

“わが国”，“その後”などのような用語を特殊辞書として持てば分かち書きができる。これをすべて網羅することはなかなか難かしいが，使われる頻度の高いものを抽出し分かち書きに利用することができる。

6）用語の一部が平仮名書きされている場合

例　か動，ひん度，赤ん坊，丸の内，風どう

このような用語は人手による整理以外に方法はない。しかし，一つの分野に限れば発生する用語が定まってくるので，これを集めテーブルにして処理することができる。

7）助数詞が漢字列に付いた場合

例　9月末東京で，11月紹介した

このような場合のため，数詞，助数詞についてはJICSTファイルより分析中である。

特殊用語，分かち書きをしないように制限する用語をテーブルに持つが，対象分野によって追加，削除を行わなければならない。

たとえば，次のような例もある

例　1）同年着工した工事は　　2）同年次の人が

この場合，“同年”で使われるケースが多いため，特殊用語として使うと，2）の“同年次”というような用語が出てくる。これは“年次”に同が付いたもので“同年”に付いたものではない。

このように一つの用語を特殊用語として分離に用いると，必ずといって，その例外が現われる。このため，これらは人手によって補正しなければならない。

3－2　漢字列の中で分かち書きが行い難い場合

1）助詞を省略している場合

例　(i)　「本会議で起草続ける。」　“起草を続ける。”と書かれる“を”が省かれている。

(ii)　「定年近い下級公務員の悩み。」“定年が近い”，“定年の近い”と書かれる“が”，または“の”が省かれている。

2），．が省かれている場合

例「マーケットについて説明開発製品の育て方と …… 」“，”が省かれているため“説明開発製品”という漢字が取り出されている。

3）用語の使い方にあいまいさがある場合

例　彼は毎日新聞を読む。

これは，“毎日”で切れるのか“毎日新聞”で切れるのかわからぬため発生する問題である。

4）漢数字を算用数字で書いている場合

例　1次元，1元的

これは，誤りとも取れないし表記のゆらぎとして考えなければならないものである。

5）カタカナと漢字が結びついて用語を作っている場合

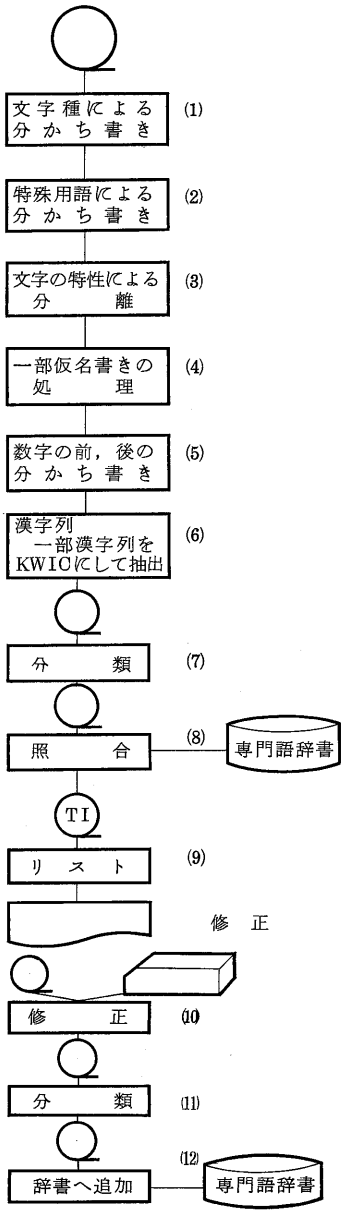
例　メートル法，フランス語

外来語でカタカナ表記した文字列と漢字列の結びつきは強い。これについては今後の研究テーマの１つである。

１）～５）については人手による分かち書きを行うか，もっと本格的辞書を準備し，品詞付け，統合処理まで行わなければ分かち書きが行えない。これらについての本格的研究は別の機会に発表したいと考えている。ここでは例のみを示す。

漢字構成の用語収集プロセス

漢字列のいろいろな問題点を調べてきたが，この経験をプロセスにまとめてみると次のようになる。



(1) 文字種による機械的分割を行う。

(2) "上記"，"その後" …… 等の特殊文字列の分割を行う。

(3) 漢字列の末尾から，上，下，中，前 …… と尚，或，且 …… 等の文字の分離

(4) 赤ん坊，か動 …… 等の一部仮名書き用語の分かち書きを行う。

(5) 数字の前後の分かち書き

(6) 漢字列，一部漢字列をＫＷＩＣにして抽出

(7) 同一文字列のＳＯＲＴと集約

(8) 専門語辞書との照合
該当するものにマーク付けをする。

(9) 専門語辞書と一致しないもののみをリストする。

(10) 漢字列ファイルを修正する。

(11) 同一文字列にＳＯＲＴし集約する。

(12) 専門語辞書へ用語を追加する
使用頻度を加え，累積使用頻度を作成する。

このようにしてできた専門用語ファイルに読み仮名を付け，訳語を付け，さらに充実したファイルにしたい。このように統一的な処理プロセスで多くの用語を組織的に集めるべきであろう。このようにして集めた用語ファイルは，次のような応用分野が考えられる。

1）機械翻訳の用語ファイル　　3）カナ漢字変換用辞書ファイル

2）自動分かち書きのファイル

などとして使える。これらを利用して日本文の校正作業とか，言語の基礎的研究，機械翻訳のシステム等へ発展さすことができる。

5. 結　　語

JICST 抄録ファイルより25万件の漢字列を抽出し，用語を取り出す上での問題点，解決方法，その処理プロセスを得た。また，この処理に必要なテーブルも作成することができた。

正確には調べられていないが5万種の漢字列は約4万数千件の漢字用語になる。

今まで長単位の漢字列は取り扱いにくいものと考えられてきたが，この分析を通して一つ一つ問題を解決すると，さほど困難なものでないことがわかる。

5－1　今後の問題点

長単位の漢字用語を収集することを前提に分析を進めると次のような問題点があることがわかった。これらは今後の多くの人々によって解明されることを期待したい。

1）長単位の用語を収集すると用語の数が無限に増えるのではないかという不安がある。

しかし“経営”とか“電気”などという一つの分野を限れば用語の数はあまり増大しない。一つの分野の漢字の専門用語は数十万語程度であろう。

2）漢字列は調査対象件数を増加させるとそれにつれて増えるが，これは二つの要因がある。一つは“上記”“その他”“この他”のように，たまたま他の漢字列と結びついて漢字列を作るもので，これは調査件数の増加に従って漢字列と結びついて漢字列を作るもので，これは調査件数の増加に従って漢字列が増加する。

他の1つの要因は新しい専門用語の増加である。これはある語数（一つの分野で数十万語）を越えると増加率は減少する。

これはまだ仮定であるため，今後さらに分析したい。

3）日本語の概念と表現が他の言語の概念と表現上，どの様な場合に一致し，どのような場合に一致しないか，を研究する。これは長単位の用語をさらに分かち書きするために必要である。

謝　辞

本稿を終えるにあたって，日頃，いろいろと助言をいただいている日本科学技術情報センターの中井浩氏，茨城大学石綿敏雄教授，京都大学長尾教授に感謝する。

添　付　資　料

1. これは漢字列の中で分かち書きが比較的行いやすい場合①，②，⑤のテーブルである。

或いは，敢えて，洗いざらい，当り，当たり，或は，有難う，案外，ある程度，以下，以上，一方法，一方的，一応，以然，一部分，一般化，一般用，一般的，一般，以内，以外，以後，一向，生き生き，今まで，何れ，今さら，一斉に，勢い，一段と，一体，一層，一段，末だ，一円，痛く，一方側，一方，一昼夜，一旦，一度，今迄，今に，一杯，一様，一番，一度に，生きいき，如何，幾つか，旧，一方向，以前，以来，以降，一時点，一時間，一時的，一時，一切，今一度，依然，幾分，色々，一役，内訳，延々，鋭意，大幅，大きな，大いなる，住々，

（以下省略）

2. この表は，尚，或，且，毎，又，迄，蓋，為，及，第，目，間，約，例が漢字例の末尾にきた場合これらの文字を分離することを制限するためのテーブルである。

該　（当該）

為　（両為，不為，念為）

及　（言及，普及，追及，遡及，過不及，波及，可及的，論及，聞及）

第　（落第，次第，面目次第，手当次第，勝手次第）
（数字）

等　（対等，高等，優等，同等，中等，下等，均等，事等，何等，彼等，恒等，不均等，動等，劣等，差等，初等，数等，平等，悪平等，不平等）

（以下省略）

3. 数字の前にくる文字列

月　産	毎　日	年間約	等
約	月　商	数	平　均
人　口	蔵　書	紀元前	計
年　間	年　収	過　去	

（以下省略）

4. 数字の後にくる文字列

～アール	～箇月	～月時	～カ所
～安打	～カ条	～月時点	～カップ
～委員会	～塊り	～月後	～カラット
～吋	～カ年	～月号	～眼
～インチ	～株	～月末	～回戦
～因	～巻	～月目	～カ月ぶり

（以下省略）