自然言語処理 41－4
（1984. 1. 26）

語と語の関係について

田中康仁　　水谷静夫　　吉田　将
姫路短期大学　　東京女子大学　　九州大学

1 はじめに

この研究は語と語の関係を基にした自然言語の知識データの収集方法について述べる。仮名漢字変換システムの同音異義語の判別や機械翻訳システムの多義語の判別の基礎となる語と語の関係の知識データを多量の日本語文の中から抽出する方法と実験結果を述べる。語と語の関係の知識データが同音異義語の判別や多義語の判別には必要であることが多くの研究で述べられているが，これらを実証する知識データについては手作業によって集めざるを得ないという状況である。筆者らはこれら知識データを日本科学技術情報センターの管理システム技術編1年分のファイルから自動的に抽出し，加工し，辞書にまとめた。

2 語と語の関係についてのこれまでの研究

語と語の関係についてのこれまでの研究としては代表的なものとして次の3つを掲げることができる。

(1) 九州大学／九州芸術工科大学　吉田（筆書）／稲永等は辞書の例文，新聞・雑誌より語と語の関係のデータを約10万組程度集めている。これは手作業によって集め入力したものであり，長い年月をかけ作りあげたものである。この研究もこれを一層発展させる考えで独自の方法を作りだした。

(2) 京都大学　梅本堯夫等によって，連想による方法を用いて語と語の研究がなされている。
約1,000人の大学生に単語を提示し，その語によって連想する語を記入させるという方法を用い約300語について実験を行っている。この研究は心理学の立場から語の連想を調べたものである。
また国立国語研究所で幼児の語の連想について調べている。
言葉は人間の精神活動の表現であるから，心理学的立場からの研究は重要である。

(3) 東京大学の荻野は約20語の語彙をマトリックス表示し，その間に関係があるか否かを網羅的に調べている。
筆者（田中）も同一方法で多量のデータを処理することを考えたが作業量が膨大になり途中で中断した。例えばn個の語彙を調べるとするとn（n－1）個の関係を調べなければならない。
nを日本語の基礎語彙3,000語を選んだとしても約900万組の関係を調べなければならない。また多くの作業者でこの調査をするとしても品質のむらが発生し，労力のわりには良い方法とは言えない。関係があるという程度の強弱については全くつかめない。

(1),(2),(3)と今までの研究を調べ別の方法で語と語の関係についての知識データを大量に集めることはできないか，判断作業を少なくし，自動的に集める方法はないものかと考えた。

3 語と語の関係の知識データを集める新しい方法の提案

日本語文の解析を通して漢字列は次のような特徴を持っている。

1 1文字漢字列は動詞，形容詞，接続詞等の語幹や，一部平仮名書きした用語として用いられている。

2 2文字，3文字漢字列は日本語の基礎的概念の語彙と基礎的概念の語彙に接頭語，接尾語が接

続した派生語である。

3　4文字，5文字漢字列は慣用表現や基礎的概念語が結合してできた複合語である。

4　6文字以上の漢字列はその他の複雑な語の組合せによってできている。

ここでは，3の4文字，5文字に注目する。特に4文字漢字列を研究対象として取り上げる。これには次のような例がある。

例　早期妥結 → 早期に妥結する。　番組編成 → 番組を編成する。

4文字漢字列の約9割は2文字漢字列と2文字漢字列に分割することができる。これは国立国語研究所の野村によって調べられている。また，この分割された2つの用語は助詞や一定の語尾表現を用い文章にすることができる。稲永／吉田（筆者），水谷（筆者）等の研究をまとめ，さらに発展させた。この表現形式を調べると幾つかの種類になることがわかる。

4　日本科学技術情報センターの抄録ファイルによる実験

4.1　漢字列の長さとデータ件数

実験データとして管理システム技術編VOL 11 No.1～12 の1年分の抄録ファイルを使用した。この抄録ファイルの中にどれだけの漢字列があるか，また漢字列の長さと種類，延データ件数を調べた。この結果は表1，図1に示す。4文字漢字列は延べ78,304件，種類で31,932種類であった。

文字数	種　類	延 件 数
1	1,230 (種)	122,114 (件)
2	10,095	295,301
3	15,690	69,066
4	31,932	78,304
5	15,681	26,356
6	11,508	15,499
7	4,959	6,129
8	2,398	2,768
9	1,100	1,264
10	505	556
その他	475	514
計	95,573 (種)	617,871 (件)

表　1

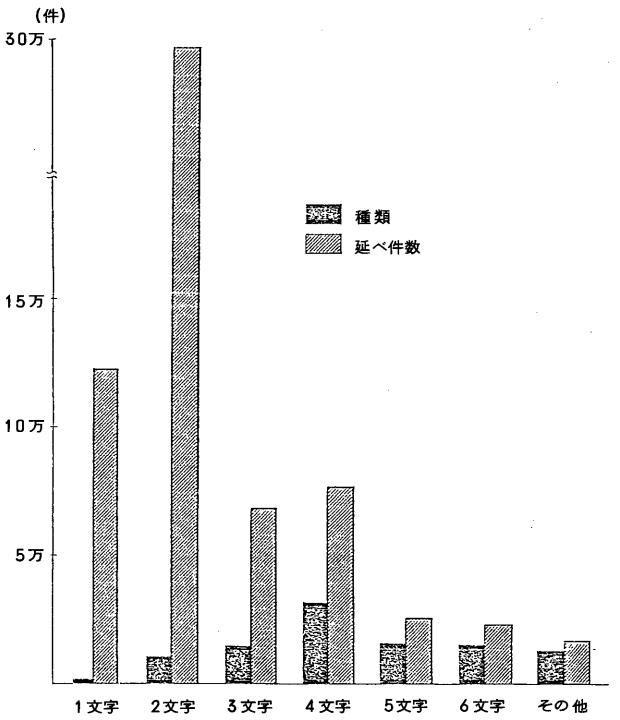


図1　漢字列の長さとデータ件数

4.2　4文字漢字列の発生度合

4文字漢字列の延べ件数，種類は明らかになったが，はたしてこれら漢字列は今後どの程度増えてゆくものか，管理システム技術編の中の4文字漢字列全体の何%程度を網羅しているかという疑問が起る。このために5千件毎に新しい漢字列がどの程度増加するか等について調査した。表2が調査結果である。この表からわかるように約3万種類の4文字漢字列で約7割程度を占めていることがわかる。この表から推測するとあと2～3年分程度のデータを分割すれば9割以上の4文字漢字列を占めることがわかる。これは管理システム技術編に限定したためにこのような結果が得られたのである。

4文字漢字列データ件数	4文字漢字列の種類	5,000件毎の新しい漢字列	新しい漢字列の発生度合
0～ 5,000	3,345	3,345	66.90
～10,000	6,102	2,757	55.10
～15,000	8,647	2,545	50.90
～20,000	10,996	2,349	47.00
～25,000	13,108	2,112	42.20
～30,000	15,073	1,965	39.30
～35,000	17,014	1,941	38.82
～40,000	18,932	1,918	38.36
～45,000	20,737	1,805	36.10
～50,000	22,618	1,881	37.62
～55,000	24,363	1,745	34.90
～60,000	25,986	1,626	32.52
～65,000	27,597	1,611	32.22
～70,000	29,219	1,622	32.44
～75,000	30,793	1,574	31.48

表2　4文字漢字列の種類と増加数発生度合

4.3　語と語の関係についての分析

4文字漢字列31,932種のデータのうち，2語と2語に分割できないものや不適当な漢字列約4,000件を取り除き28,033種類のデータを知識データとして集めた。

さらに，これら漢字列を構成している要素の語類を次のように定めた。

T（term）体言，P（predicate）Pv，Paに細分「～スル，～サレル」がPv，「～（的）ダ／ナ，～トシテイル」がPa両者可能ならPvとする。M（modifier）連用修飾的「～的ニ」に言い替えられれば，ここに入れる。　F（affix）接辞的 例として ～自体，～相互等に分類し，助詞，語尾の種類によって語と語の関係を46のカテゴリーに分類した。その中の幾つかを示す。

(1)　MにPv　　迅速・処理
(2)　Tが／をPv　　会長・辞任
(3)　TによりPv　　写真・判定
(4)　TにおけるT　　米国・犯罪

これらの詳細な分類区分とデータの種類，延件数は添付資料1，2を参照されたい。分類の形式と種類の多いものを20ケあげてみると表3のようになる。

順位	区分コード	分割の形式	種類
1.	55	Tニ対スル／関スルT	4,411
2.	41.1	Pvスル／サレルT	3,911
3.	30	PaナルT	2,844
4.	42	Pvスル為ノT	1,947
5.	41.0	Pvスル／サレルT	1,254
6.	21.0	TガPv	869
7.	04	PvシテPv	818
8.	50	TノT	702
9.	56	TニオケルT	684
10.	02	Mニ／トPv	658
11.	51.1	TナルT	577
12.	57	Tノ為ノT	496
13.	00	TF	479
14.	53.0	Tガ有スルT	449
15.	10	TガPa	373
16.	52.1	TノタメノT	310
17.	41.2	PvシテノT	302
18.	58.2	TニヨルT	294
19.	26.1	TニヨリPv	273
20.	43	Pvスル／サレルT	250

表3　語と語の関係，種類の頻度順

4.4　4文字漢字列の分析手順

4文字漢字列の分析図2の手順に従って行った。

手作業での分析では図3のカードを用いて

行った。分析カードは今回は手作業で作成したが電子計算機による出力結果を用いればさらに合理化される。漢字列を電子計算機のファイルとして管理すれば再入力の無駄を省くことができる。しかし，漢字データの中には１％程度の誤りがあるので，これを除去しなければならない。仮名付け，区分コード付けはなんらかの方法で人手を介さなければならない。このように少しの工夫により多量の語と語の関係による知識データを安く，早く作成することができる。

この方式の特徴をまとめてみると次のようになる。

(1) 特別の専門的知識を必要としない。

(2) 機械的操作により知識データを集めることができる。

(3) 処理回数が増加するごとに処理量は減少する。

この処理手順で得られた結果は約250頁の資料としてまとめた。

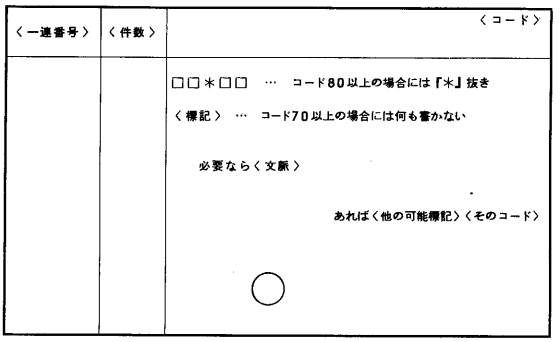


図３　カードの記入形式

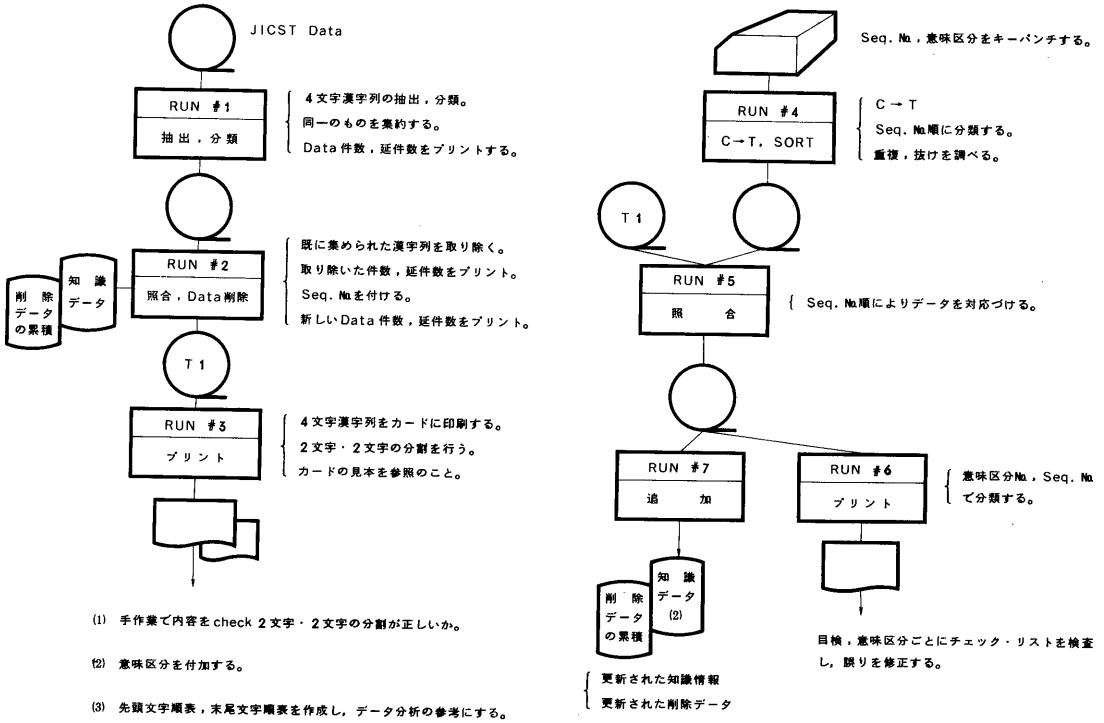


図２　４文字漢字列の処理プロセス

5. 語と語の関係による知識データの利用分野

4までに述べた方法により収集した知識データは次のような分野に利用することが出来る。

(1) 仮名漢字変換システムの同音異義語の判別に利用する。

例1　ジシン ヲ ソウシツ スル。

① 自身を喪失する。
② 自信
③ 地震
④ 時針
⑤ 侍臣
⑥ 磁針

知識データの中に自信と喪失があるので，②が選ばれ「自信を喪失する」という変換結果が出力される。頻度情報を用いれば他のものが最初に出力されているかもしれないが知識データにより正しい選択がなされる。

例2　キケン ヲ　ボウシ スル。

① 危険を ⑤ 防止する。
② 気圏 ⑥ 紡糸
③ 棄権 ⑦ 暴死
④ 貴顕 ⑧ 眸子
⑨ 帽子

サ変動詞から⑤防止と⑦暴死に限定される。
さらに知識データから

{危険を防止する}
{棄権を防止する}

になる。これ以上の判別は2項関係の知識データでは判別できない。3項以上の関係の知識データに頼らざるを得ない。このような場合は非常に少なく同音異義語のほとんどは2項関係の知識データにより判別できる。

(2) 機械翻訳の多義語の選択

例　協約の期間

協約　＝ an agreement, a convention, a pact
期間　＝ a term, a period of time, a life
協約・期間 ＝ a life of an agreement

協約・期間 ＝ a life of an agreement があればこの訳語が選ばれる。
機械翻訳の多義語の問題については今後の研究に期待していただきたい。

(3) 手書文字認識，音声認識システムに利用する。

手書文字認識では認識の技術を向上させることが最重要であるが，第二位になる文字の中にも正しい文字があらわれる。このため知識データを利用し，文章としての正当性を検討するために使用する。この研究は筑波大学と協同で研究を進める予定である。音声認識システムにも同様の技術が利用できる。

(4) 自然言語の解析を支援するデータとなる。

日本語文を解析する際に，パーサと文法規則（学校文法）程度だけでは解析木が急増することがある。この急増を押さえるためには語と語お関係の個別規則を導入しなければならない。このために語と語の関係の知識データが使われる。

6. 今後の課題

今回の分析は知識データの収集方法の開発とカテゴリー別に分類することが主目的であった。今後に残された課題をまとめてみると次のようになる。

(1) 管理システム技術編を数年分分析する。

(2) 多くの分野から知識データを集める。

日本科学技術情報センターのファイルには次の12分野がある。

(i) 管理・システム技術	(ii) 電気工学	(iii) 化学・化学工学
(iv) 環境公害	(v) 機械工学	(vi) 原子力工学
(vii) 物理・応用物理	(viii) 金属・鉱山・地球科学	(ix) 土木・建築工学
(x) 生命化学	(xi) エネルギー	(xii) 化学・化学工業（外国編）

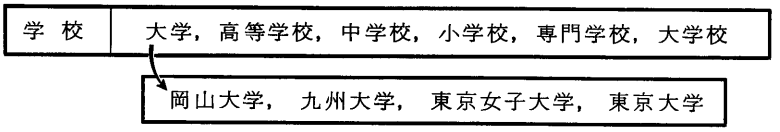
等の各編がある。この中から知識データを抽出すると仮定すると，各データは共通部分と各分野固有のデータから成り立っている。共通部分を0.4，固有の部分を0.6とするとして12の分野を調査する。1ケの分野で平均2.8万件の語と語の知識データが得られるとする。

$2.8 \times (0.4 + 0.6 \times 12) = 21.28$　約21万件の知識データが得られる予定である。

(3) 長単位用語の語と語の関係のための意味の小集合を作る。

例　学校における教育

この場合，学校には次のような語が関係する。



意味の集合には上位，下位の関係を作らなければならない。

(4) 5文字漢字列，3文字漢字列の分析

4文字漢字列の分析を行ってみたが，これで全ての知識データの収集が完了したわけではない。5文字漢字列の分析，3文字漢字列の分析を行い知識データを増やしてゆきたい。

(5) 今までに集めた知識データの各カテゴリー別の分析

今までに集めた知識データをさらに分析し，精度を上げると同時に多量のデータ分析のために自動収集方法，自動カテゴリー化等を考える。

(6) 知識データの応用

知識データの応用について幾つかの分野をあげているが，今後はそれらを実験すると同時に他の応用分野をみつける。

7. おわりに

この研究は漢字列，特に4文字漢字列により知識データを収集する方法の開発であった。この方法は機械的操作を多く取り入れているので，知識データを多量に，早く，安く，容易に収集することができるのが，特徴である。

今後は仮名漢字変換システムや機械翻訳システムに組込んで実用化したいと考えている。

最後に，この研究に協力して下さったＪＩＳＴの 長田孝治氏，日本科学技術情報センターの 中井浩氏，佐藤雅之氏に深く感謝する。

参考分献

(1) 梅本堯夫　連想基準表 ―大学生1000人の自由連想による― 東京大学出版会 1969.3

(2) 野村雅昭　四字漢語の構造　国立国語研究所報告54「電子計算機による国語研究Ⅶ」1975.3

(3) 野村雅昭　現代漢語の語構成について　情報管理 VOL 18 No. 11　1976. 2

(4) 稲永，小西　カナ文字文の機械処理のための辞について　AL 76-39　電子通信学会技術研究会資料　1976

(5) 石綿敏雄　コンピュータと言語　日本語別巻　岩波書店　1978. 3

(6) 水谷静夫　語構成概説　岩波国語辞典第三版　岩波書書　1980. 3

(7) 荻野綱男　名詞と動詞の結合の数理的取り扱い「ソフトウェア文書のための日本語処理の研究」　情報処理振興事業協会　1982. 3

(8) 山梨正明　意味と知識構造　言語学からの意味表現モデルの理論的検討　数理科学 No. 240　1983. 6

(9) 田中康仁，水谷静夫，吉田　将　語と語の関係について　情報処理学会　第27回全国大会論文集　情報処理学会　1983. 10

(10) 田中康仁，水谷静夫，吉田　将　語と語の関係について ―知識データ収集方法についての試み―　第20回情報科学技術研究集会論文集　日本科学技術情報センター　1983. 10

添付資料 1

コード	標記	例	意味的規定 併せて 備考	種類	延件数
00	TF	数値自体 団体相互		479	844
01	Mニ／トPa	一様単調 完全自動		22	29
02	Mニ／トPv	迅速処理 積極介入		658	1,535
03	PaデPa	豪華美麗 温厚朴訥		55	117
04	PvシテPv	抽出整理 神出鬼没		818	2,072
05	Tヤ／モT	科学技術 老若男女		177	599
06	MカツM	縦横無尽 迅速円滑		12	14
10	TガPa	容姿端正 健康無類	TがPaであること／さま 「威風堂々」もここに入れる	373	736
14	TニPa	学校至近 機械特有	Tに対しxがPaであること／さま	18	22
15	TデPa	業界一位 南半最大	TでPaであること／さま	40	62
21.0	TガPv	株価騰貴 職住接近	TがPvすること Pvは他動詞的でもよい なおcf. 22.	869	2,354
21.5	〃	疾風迅来	TがPvするようにxがPvすること	0	0
22	Tガ/ヲPv	権威失墜 会長辞任	Tが/をPvすること この場合のPvはヲ格を取っても自動詞的	108	367
23.0	TヲPv	統制撤廃 番組編成	TをPvすること cf. 25.	4,289	13,454
23.5	〃		TをPvするようにxをPvすること	0	0
24.0	TニPv	憲法違反 因数分解	TにPvすること Tが所や時なら24.3	187	316
24.3	〃	国外追放 早期妥結	所／時にPvすること	235	402
24.5	〃	暗中模索	TにPvするようにxにPvすること	1	1
25	Tデ／カラ／ヲPv	中途退学 空中撒布	位置（多くは場所的）Tで／から／をPvすること ヲの例は「戦線離脱」	229	425
26.1	TニヨリPv	写真判定 実物教育	Tを使ってPvすること	273	739
26.2	〃	傷害致死 結核死亡	Tが原因でPvすること	13	17
27	TトPv	父兄面談 海外貿易	Tを相手／と共にPvすること 「親子対面」→ 10, 「記者会見」→ 24.0	3	3
28	TトシテPv	団体交渉 個人所有	Tの資格・状態でPvすること	42	112
30	PaナルT	緊急事態 巨大粒子	TがPaであるそのT	2,844	5,352
41.0	Pvスル／サレルT	適応能力 違反行為	TがPvする／されるそのT	1,254	2,496
41.1	〃	捜査過程 生産意欲	x（≒T）がPvする（ことに結び付く）T 42に移せるものはそちらに移せ	3,911	9,570
41.2	PvシテノT	著作収入 移動回数	x（≒T）がPvすることから生ずるT	302	1,009
42	Pvスル為ノT	印刷方式 冷凍時間	そのTを使ってPvするT	1,947	5,577
43	Pvスル／サレルT	爆発地点 許可年月	その所/時Tで Pvする／されるT	250	559

添付資料 2

コード	標記	例	意味的規定 併せて 備考	種類	延件数
50	TのT	中小企業 日常業務	下記のどれにも該当しにくく，しかも「αノβ」で意味が一往通ずるもの	702	1,857
51.1	TナルT	非鉄金属 内縁関係	$T_{左}$である$T_{右}$　$T_{左}$が$T_{右}$の着眼点であれば，52.2 cf. 52.2の備考	577	1,426
51.3	Tガ特色ノT	機械時代 住宅地域	$T_{左}$の存在を特色とする所／時$T_{右}$	70	150
51.5	TナルT	東洋哲学 資本主義	$T_{左}$にゆかりのある$T_{右}$	89	178
51.6	T ニ似タT	米国方式 積木方式	形状が$T_{左}$に似た$T_{右}$ 「葡萄状球菌」の「状」が無ければ，ここの例となる	13	21
52.1	Tノ名ノT	第二工場 日本海流	$T_{左}$を識別呼称とする$T_{右}$	310	502
52.2	TトイウT	因果関係 会計分野	$T_{左}$の観点で考へる$T_{右}$ $T_{左}$が$T_{右}$の属性でも$T_{右}$の対象の属性でもない場合	167	425
53.0	Tガ有スルT	国家資産 郵政大臣	$T_{左}$が$T_{右}$を有するその$T_{右}$	449	1,060
53.5	〃	広告生命	$T_{左}$の有する$T_{右}$に似たもの	1	1
54	Tヲ有スルT	悪臭物質 実績工数	$T_{右}$が$T_{左}$を有するその$T_{右}$	39	89
55	T ニ対スル/ニ関スルT	位置母数 自然現象	$T_{左}$を対象とする／に関する$T_{右}$ 「ニ対スル」「ニ関スル」双方可なら「ニ対スル」で記述	4,411	12,299
55.9	T ニ関スルT	建設業者 株式課長		161	436
56	T ニオケルT	米国犯罪 未来社会	所・時$T_{左}$に実現する／している／した$T_{右}$	684	1,768
57	Tノ為ノT	軍事予算 安全装置	$T_{左}$の為に使う$T_{右}$ 「高速道路」「生命保険」etc. もここに入れる 42の混入を避けよ	496	1,176
58.1	T ニヨルT	羊毛織物 亜鉛合金	$T_{左}$を材料として作る$T_{右}$	25	96
58.2	〃	蒸気機関 移動平均	$T_{左}$の利用を特色とする$T_{右}$ $T_{左}$が用言的なものも当面はここに入れる	294	1,184
58.3	〃	台風災害 価格戦争	$T_{左}$が原因で生ずる$T_{右}$ $T_{左}$が用言的なものも当面はここに入れる	139	297
70			上記のどれに分類しても不具合なもの	44	50
				28,033	71,798

附記 (1) 言い替えにおけるテンスは無視する。受身もなるべくは採用しない。

(2) 「積極介入」を M(的) ニPv とし，「郵政大臣」を T(名) ガ有スルT とする程度の，自然さを確保する補入はしてよい。

(3) 二つ以上の解し方が出来ると思えば，最も適切と思う所に分類し，他の可能性を下部に記せよ。

添付資料 3

区分NO 230
分割の形式
T ヲ Pv
種　　類 4289種
延データ13,454件

1	圧力・測定	1
2	圧力・損失	4
3	安全・維持	1
4	安全・管理	52
5	安全・監査	1
6	安全・確認	2
7	安全・確保	12
8	安全・記録	1
9	安全・規制	1
10	安全・教育	15
11	安全・強化	1
12	安全・計画	5
13	安全・検査	2
14	安全・指導	2
15	安全・重視	1
16	安全・推進	4
17	安全・設計	2
18	安全・点検	4
19	安全・評価	1
20	安全・保障	2
21	安定・維持	1
22	安定・確保	3
23	悪臭・除去	1
24	悪臭・防止	1
25	一国・救済	1
26	一般・教育	3
27	一品・生産	1
28	一部・改正	1
29	一部・公表	1
30	一部・支払	1
31	一部・自給	1
32	一部・実用	1
33	一部・制限	1
34	引用・分析	1
35	以下・連載	4
36	印刷・開発	1
37	印刷・制御	1
38	印字・構成	1
39	印象・形成	1
40	館内・案内	1
41	因子・分析	32
42	位置・選定	1
43	医学・管理	1
44	医薬・悪用	1
45	医薬・実験	1
46	医療・分類	1
47	育成・強化	1
48	為替・換算	2
49	為替・管理	2
50	為替・処理	1
51	為替・制限	1
52	移転・見学	1
53	移動・記録	1
54	移動・研究	2
55	移動・準備	1
56	移動・推行	1
57	移動・分析	1
58	異音・解析	1
59	異議・申立	2
60	異物・検査	2
61	違反・予防	1
62	違法・摘発	1
63	飲料・生産	1
64	意見・区分	2
65	意見・具申	3
66	意見・形成	1
67	意見・決定	2
68	意見・交換	3
69	意見・差控	3
70	意見・収集	1
71	意見・調査	3
72	意見・聴収	1
73	意見・表明	1
74	意向・調査	2
75	意匠・公報	1
76	意匠・出願	1
77	意匠・審査	2
78	意匠・分類	1
79	意匠・利用	1
80	意志・決定	50
81	意志・伝達	2
82	意志・表示	1
83	意思・決定	412
84	意志・表示	1
85	意識・改革	2
86	意識・改善	1
87	意識・革新	3
88	意識・管理	1
89	意識・調査	8
90	意識・変革	2
91	意味・解釈	1
92	意味・解析	1
93	意味・記憶	1
94	意味・構成	1
95	意味・処理	2
96	意味・識別	1
97	意味・説明	1
98	意味・分析	6
99	意味・変換	1
100	緯度・表示	1
101	運営・維持	1
102	運営・監査	2
103	運営・研究	1
104	運営・継続	1
105	運営・評価	2
106	運営・分析	2
107	運行・管理	1
108	運行・研究	1
109	運行・指令	1
110	運航・研究	1
111	運航・指令	1
112	運賃・計算	2
113	運転・開始	6
114	運転・解析	1
115	運転・休止	1
116	運転・実習	1
117	運転・試験	1
118	運転・報告	1
119	運動・記録	1
120	運動・計算	1

収集した知識データの一部