自然言語処理 33－4
（1982. 10. 22）

同音語の判別

甲野 洋（国立国語研究所）

1. 目的

日本語文章の入力手段としてのカナ漢字変換は 入力方式により、①漢字一字をカナ連糸で呼び出す方法，②単語分かち書き入力，③かなべた書き入力の三種類に分けることができる。その長短を比較すると，①の方法は処理が簡単でCPU-TIMEも短いが変換率が悪く，選択時間がかかりすぎること，③の方法は将来性はあるが現時点では変換率があまりよくなく，かつ，処理に要する辞書が完全ではないことがあり，現時点では②の方法の精度をあげることが現実的であると思われる。

さて，単語単位での変換においては，以下に述べるように約20～40％といわれる同音語の判別が問題となる。同音語の判別のために現在用いられている有効な手段は頻度情報であるが，これとて完全なものではもちろんない。関連語情報・意味情報の利用は，それ用の辞書の作成が簡単ではないことと，したがってその有効性についてのデータがないこと等が問題であったのか，市販のワードプロセッサでの利用を知らない。

ところで，単語の認定には前後一語の環境があれば人間には可能であると言われている。これは，複合語についてはなおさら，単純語においてもそのかかり受けの距離が，多くの場合二語の範囲にあるためだと思われる。

そこで，本報告は，同音語の判別のために，その直前または直後の語の出現状況と，それを利用した場合の同音語の判別精度について述べる。

2. 同じ語か異なる語かの問題

カナ漢字変換は出力文の漢字含有率が多ければ多いほどよいと考えられているようだが，それは間違いだと思われる。出力は，たとえ漢字が少なくても，調和のとれた，読みやすい漢字かな混り文であるべきである。

事実，高校教科書（日本史）ではある一語が次のような形で用いられる。

あう　合い(1)，あい(1)，合う(1)，あう(1)，会っ(1)，合っ(1)，あっ(3)，合わ(6)，あわ(6)

すなわち，表記という面では「あう，合う，会う」の三種類が混在するのである。これは，表記法が統一されていないためではなく，文脈による意味の強調のしかたによる使い分けのためだと思われる。

ところで，複数の単語を，使用される形と読み，およびそれが同語か異語かの違いで分類すると次の8種類に分けられる。

1.	同語	同音	同形		
2.	同語	同音	異形	表記の違いなど	会う，あう，合うなど
3.	同語	異音	同形	発音の違いなど	「雨」＝「あめ・あま・さめ」など
4.	同語	異音	異形	活用の違いなど	会わ，会い，会う，あわ，あいなど
5.	異語	同音	同形	品詞性の違いなど	アル家，家がアルなど
6.	異語	同音	異形		核，各，欠く，書く，画く，描く

7．異語　異音　同形　　「工夫」＝「こうふ・くふう」

8．異語　異音　異形

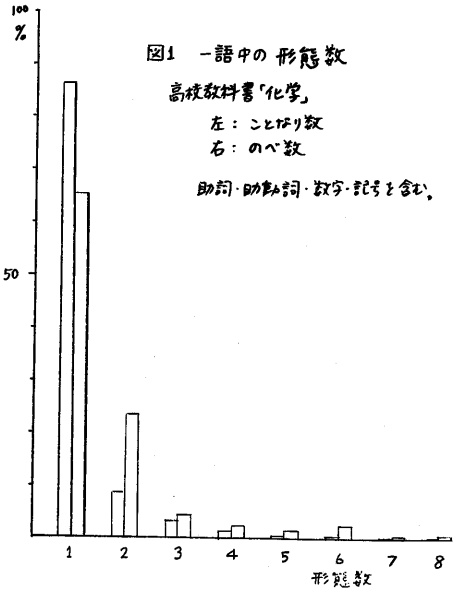


図1　一語中の形態数
高校教科書「化学」
左：ことなり数
右：のべ数
助詞・助動詞・数字・記号を含む。

異なり語数の最も少なかった「化学」において，ある一語が何語の形態（語形）を有したか（上例「あう」の場合，異なり9，延べ21）を示すと図1のようになる。すなわち，全体のうち延べで約63%，異なりで約86%が一形態しか持たず，したがって少なくとも残りの，延べで27%，異なりで14%が同じ語か異なる語かの判別が必要だったわけである。

言語処理において単語の認定は少なくとも上記の区別がつけられることが一応の目標となる。

3．同音語の種類と量

日本語は，その音節構造が単純なこと，表意文字としての漢字によって作られた語（漢語）の占める割合が大きいことにより同音語が多い。

田中章夫（文献1）によれば，同音語には次の種類がある。（　）内は新聞1紙1年分1/60の用例数）

①表記にゆれのある短単位　（例）　飲び（6）／喜び（17）

②送りがなにゆれのある短単位　　異なる（7）／異る（5）

③いわゆる同音異義語　　市立（20）／私立（17），科学（117）／化学（116）

④同音類義語　　機具（1）／器具（13），追求（15）／追及（38）／追究（2）

これらのうち，機械で判別しなければならないのは③の同音異義語であって，①②は表記する者の好みによるし，④の同音類義語の判別は無理だろう。③の中も，その使用度によってともに一般語か非一般語を含むか，品詞性が同じか，慣用的用法があるか，アクセントは同じかによって分類することもできる。（文献2）

計算機処理の場合，普通には考えられなかった，次のような同音異義語セットも示されている。（文献1）

①活用語の変化形を含む同音語セット　　長さ（20）／流さ（2）
　良く（6）／欲（5）／翼（7）／翌（13）／浴（172）

②現代かなづかい表記で同一よみがなとなる同音語セット
　小売り（16）／公吏（1），王（80）／追う（8）／負う（5）

③固有名詞を含む同音語セット　　佐藤（276）／砂糖（17）

④助詞・助動詞・接辞などを含む同音語セット　　まい（65）／毎（13）／枚（124）
　たく（29）／宅（44）／卓（3）

これらの判別は，その例数も多いため避けて通れない処理である。

同音語の量は，例えば「新明解国語辞典」を調べるとその36.4%が同音語であるという（文献3）。これを新聞語彙調査（1紙1年分1/60）で調べると次のようになる。

新聞（1紙1年分 1/60）に現われた同音語

	同音語セット	延べ語数	出現率	用言の量	かな書き語の量
全体	4141	67037	16.8%	1.7%	2.4%
1万位以上	1146	59861	16.6%	1.6	2.3
5千位以上	566	47864	14.6%	1.4	2.0

（助詞・助動詞・記号等を除く）

また、同調査のデータから別に抽出し、助詞・助動詞・固有名詞・記号等すべてを含め、出現形レベルで集計すると次のようになる。

全体延べ	同音語セット	延べ語数	出現率	異形語数
623008	5334	341659	54.8%	16963

助詞・助動詞等の割合が多いため、短単位でカナ入力するとその半数以上が同音語となることがわかる。

4. 人間による同音語判別の手がかり

国立国語研究所の調査（文献2）によると、同音語は示される文脈の程度により、図2のように判別の精度があがるということである。

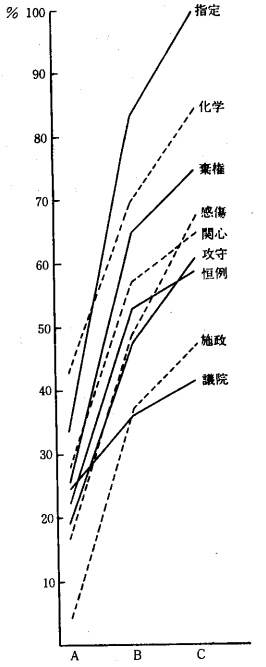
すなわち、

A　単に「シテイ」のようにカタカナ書きのことばをあげ、思い出せる漢字表記の語を書かせた場合

B　「シテイする」のように最小の文脈を与えて漢字表記させたもの

C　「シテイされた場所に、時間通りに集まる」のように意味をあきらかにする例文を与えて漢字表記させたもの

〔同音語と文脈〕



Aで、被験者（高校生・大学生計210名）は自分の知っている同音語「指定・師弟・私印」などのうち、最も一般的（使用頻度が高いことで類似できる）な語を書くと思われる。これは、頻度情報で順序づけられた同音語辞書を機械が引いて表示したのに等しいと思われる。Bの文脈は、ほぼその語の前または後（20問のうち17問が後）に判別のキーがある。

図に見る通り、A→B→Cの順に正答率が高くなっている。この調査の目的がカナ書きされた文が人間に誤りなく読まれるのにはどのような工夫が必要であるかに重点がおかれているため、直接には機械による同音語判別に利用できるわけではないが、

語結合・品詞性・慣用的用法等の判別条件と文脈や位相が、二重に働く文を作れば理解しやすい。文脈や位相だけを手がかりとする文を作ると、その手がかりに対する読者の経験・知識の有無に左右されるので誤解をまぬく可能性がある。

という結論は示唆的である。

5. 方法

上の結論は、機械による判別という立場に立てば、次のように変えることができよう。

① 複合語における語の関連情報を利用すること。
② 語の品詞性を利用すること。
③ 慣用的用法、すなわち連語情報を利用すること。
④ 文脈の情報を利用すること。
⑤ 入力文および単語に位相の情報をつけ、これを利用すること。しかし、この情報の精度によっては、かえってエラーが生じることがある。

④⑤の情報については、研究課題としては重要だが当面の利用は期待できない。
①②③の情報の利用は、次のように理解できる。すなわち、

① 学会-開催
② 指定-する
③ 風雪に-堪える

の例で考えれば、どれも同音語の直後の語が何であるかを知ることができれば、それだけで（その品詞が何であるかや、どのような意味を持つかや、日本語としての常識をどのように入れるかを考える必要がなく）、判別が可能なように思われる。しかも、この方法は、辞書を作るための大量データさえあれば、計算機の力によって実現することができる。

この方法はすでに何人もの人が提案していることであって基本的には何ら目新しいものではない。しかし、計算機容量の増加と大量データの蓄積によって本格的な実験が可能な環境にあり、それだけの方法によって、どれほどの効果があるかを知る意義はあると思われる。

6. 効率

先の新聞データを整理すると次のような結果が得られる。

同音語	語形・表記	次の語と頻度
いか	医科	だいがく(1), を(1)
	以下	、(15), いち(1), いわゆる(1), かかげる(1), かんとく(1), が(1), さんびゃく(1), じゃっかん(1), そうほう(1), ぞっかん(1), つぎつぎ(1), つみ(1), で(1), とびさき(1), どくしん(1), に(5), の(16), は(5), 。(3)
	生か	せる(1)
	行か	ない(1), なく(1), れ(3)
	いか	ず(2), ない(4), なけれ(1), なる(5), に(17), ん(4)
計	5語	延べ99語、異なり31語、同音語（異なり2，延べ27）

このとき、同音語「いか」の判別に、直後の語を用いると、原データをカナ入力した場合、「いか-ない」「いか-に」の連続の場合だけ「行か／いか」「以下／いか」の判別ができない。したがって、判別力は(99-27)／99＝0.727と計算できる。

この方法での先のデータの判別力の合計は 0.688 となる。

同音語のセット	延べ語数	異形語数	同音語の直後の語 異なり語数	そのうち 異なり	延べ
5334	341659	16963	105838	5964	106717

すなわち、同音語の直後の語の情報を用いることにより、全体の54.8%ある同音語のうち、68.8%が判別でき、結局全体の17.1%が判別不能となる。

この辞書には、助詞・助動詞が含まれており、かつその語数が非常に多い。次に示すのは上位10語であるが、ほとんどが助詞・助動詞である。

同音語	判別力(のべ)	判別力(にほり)	頻度(のべ)	異形語数	次の語(のべ)	同音語(のべ)	同音語(ことなり)
ノ	0.8923	0.9944	22189	58	6078	2389	34
ニ	0.7014	0.9688	14451	30	3566	4314	111
オ゜ →ヲ	0.9590	0.9964	10339	16	3085	423	11
ハ	0.6892	0.9914	10158	17	3172	3157	27
テ	0.7814	0.9926	8423	5	1779	1841	13
タ	0.4384	0.9684	8240	5	1584	4627	50
ガ	0.7482	0.9941	7953	13	2566	2002	15
ト	0.6558	0.9801	7740	22	1967	2664	39
シ	0.1769	0.8950	6769	50	543	5571	57
デ	0.6763	0.9904	6674	10	1886	2160	18

助詞・助動詞の認定をおこない、これを除けば大幅に判別力はあがると思われる。助詞・助動詞などを除いた同音語は、先に示したように、全体の16.8%であるから、判別力がこのままとしても 16.8% × (1 − 0.688) = 5.24% だけが判別不能として残ることになる。

カナ漢字変換の精度が単語レベルで、82.9%、または94.76%であることについての評価はさまざまだろう。ここでは問題としない。

7. 問題点

この方法について問題点が大きく二つある。一つは辞書の効果についてであり、他の一つは辞書の大きさについてである。

この辞書は、いわば日本語についての用語の常識を集約したようなものといえる。そのために従来不可能であった複合語や慣用語の処理がかなりできるようになる。しかし、たとえば、

本を大阪の書店で買った。　　国立国語研究所

の「本・国立」は直後の「大阪・国語」とは意味的に関係がないのに辞書に登録されてしまう。このような辞書が効果を発揮するわけがないという批判があろう。しかし、このように二語三語先の語にかかる用例は相対的に少なく、頻度情報によって処理が可能と思われる。

また、用例数の少ない語について、たまたま登録された語の連続が再びあらわれることは少なく、あってもその連続でないことが多いという批判がある。これに対しては、逆にあまり用いられない語は決まった形で使われることが多い、一般的には、使用率の大きい語には多義語が多く、使用率の小さい語は単義語が多い(文献7)ことが知られている。したがって、それだけ使用率の小さい語は用いられ方も単純であると予想される。

もちろん、何らかの加工をした方が効率はよいだろう。

辞書の大きさについては、問題が大きい。同音異形語 16963 に対し、次の語は 105838 と約6倍にも増える。複合語を辞書に登録する方法に比べ、同じ見出しを何度も繰り返さない分だけ、辞書は小さくなるものの、この辞書はかなり大きい。

また、現時点では、助詞・助動詞と自立語の区別をしないため、慣用句等の判別に力がない。これを区別し、次の自立語を登録した場合には次の語に対する拘束力が複合語ほどないため、辞書は大きくなる。たとえば、「自己」の直後の語「の」のその後には、「事故」のそれとは異なる語ではあるが、「出し、利益、把握、最高、認識、資質、よりどころ、著書、立場、すべて」という10語がかくれている。

あらゆる現象を記述できるという言語本来の性質のため、この種の辞書が大きくなることは避けられないだろう。これに対し、意味情報を利用する方法が検討されてはいるが、現時点では望み薄であり、本報告の趣旨に反する。

参考文献

1. 田中章夫「新聞の語彙調査の同音語と同形語」(「電子計算機による国語研究Ⅲ」1971)
2. 国立国語研究所「同音語の研究」(国研報告20, 1961)
3. 望月八十吉『中国語と日本語』(1974, 光生館)
4. 田中武美・吉田将「かな漢字変換による日本語入力システムに関する調査」(九大工学集報, 1977)
5. 牧野寛「カナ漢字変換入力法」(情報処理 Vol.23, NO.6, 1982)
6. 木村健・遠藤安彦・小椋史彦「日本語文入力用カナ漢字変換システムの試作」(情報処理 Vol.17, NO.11, 1976)
7. 国立国語研究所「現代雑誌九十種の用語用字」(国研報告25, 1964)
8. 田中康仁・岡坂良雄・長田孝治「同音異義語の解析——専門分野における」(自然言語処理32-5, 1982)

(本研究は昭和57年度文部省科学研究費特定研究「言語の標準化」(1林大班)の一部を受けて行なわれた。)