かな漢字変換における固定的共起表現

小山 泰男　安武 満佐子　吉村 賢治　首藤 公昭
（福岡大学　工学部）

1．はじめに

慣用表現は、その構成語間に強い結合関係を持ち、個々の構成語から表現全体の意味を捉えることが難しく、自然言語処理において取り扱いが重要である。かな漢字変換においては、慣用表現の変換間違いに対するユーザのストレスが大きく、慣用表現を巧く取り扱う必要がある。例えば、「腹を立てる」のような表現を1文節として変換効率を上げることが考えられるが、そうすると「（バッターボックスに）原を立てる」といった変換が不可能になるなどの弊害が現れる可能性がある。さらに「立てる」と「たてる」というユーザによる表記のゆれを考慮した操作性を考えれば、「腹を/立てる」と2文節にした方が遥かに有効であることがわかる。そこで、慣用表現を含む語の固定的な共起表現を構成語で分割した形状で網羅的に収集し、これらを構造をもった要素（単語）の列と一つの要素的な意味を持つ単位表現という二通りの構造をもって取り扱うかな漢字変換システムを試作した。以下では、本システムの特徴である種々の固定的共起表現の概要、文節の取り扱い及び派生する表記の取り扱いを示し、処理の概要を説明する。

2．語の固定的共起表現

特定の語が特定の形式で共起し、各語の通常の意味から表現全体の意味を統合・生成することが困難と思われる表現は、大きく、付属語的表現と自立語的表現に分かれ、後者はさらに一語性表現、多語性表現の2種類に分けられる。

2．1　付属語的表現（2,566語）

付属語的表現は主に概念間の関係を表す助詞的表現（関係表現）と話者の判断、態度、テンス、アスペクト等を表す助動詞的表現（助述表現）に分けられ、拡張文節[1]を構成する重要な要素である。これらは形状的に文節区切りのない付属的語表現と、文節区切りのある付属語的表現に分ける。

文節区切りのない付属的表現（576語）

拡張表現 392語	上で、終わる、回る、掛けて、損なう
複合助詞 170語	とばかり、よりか、さえも、かは、にも、のなんの
その他 14語	（走っ）てく、（飛ん）でる、（書い）とく、（噛ん）どく

文節区切りのある付属語的表現（1,990語）

関係表現 590語	さえ/して、さえ/なければ、た/ところで、に/際して
助述表現 1400語	ずに/いられない、すら/なかった/はず

2．2　一語性表現（20,015語）

一語性表現は語が隣接して結合しており他語の挿入や語順の変更が通常許されない表現である。これらは、構成語の意味から全体の意味を合成できないと思われるもので、処理上一まとまりで取り扱われる。ここでも、文節区切りのない表現と、ある表現に分ける。

文節区切りのない一語性表現（6,956語）

四文字熟語 2,495語	我田引水、不惜身命、私利私欲、叱咤激励
副詞＋に 233語	きちきちに、かっきりに、近々に、さんざんに
副詞＋と 1,136語	あたふたと、ゆっくりと、かさかさと、ふさふさと
副詞＋の 348語	いささかの、ゆるゆるの、生憎の、かさかさの
副詞＋する 1,563語	青々する、あくせくする、ゆっくりする
くる、する 218語	相手に（する）、お釣りが（くる）
その他 963語	哀れっぽい、薄べったい うら悲しい、脂っこい

文節区切りのある一語性表記（13,059語）

格言・ことわざ 2,311語	馬の/耳に/念仏、鴛鴦の/契り、売り/言葉に/買い/言葉
その他 11,748語	息を/荒くする、板挟みに/合う、一生の/不作だ

2．3　多語性表現（7,414語）

一語性表現に比べ語の隣接結合の度合は弱く、表現中に他語の挿入が許される表現である。多語性表現における語の共起の度合は千差万別であるが、構成語の意味から全体の意味を合成できないと思われるものを中心に収集した。

形容詞活用 704語	態度が/大きい、入りが/少ない
五段動詞活用 4,526語	心が/沈む、気を/吐く、油を/売る
一段動詞活用 1,920語	座を/占める、ネタが/割れる
その他 264語	世も/末（だ）、後に/引け（ない）

3．固定的共起表現を用いたかな漢字変換

かな漢字変換における文節分かち書き処理において、付属語的表現と一語性表現は1文節として取り扱われ、変換後の候補表示は表現中の文節区切りに応じて多文節として扱われる。前者を拡張文節、後者を見掛け文節とよぶ。多語性表現は係り受け処理において他の弱い共起に対して強い共起として取り扱われる。

3．1　拡張文節と見掛け文節

拡張文節は固定的共起表現の付属語的表現を付属語、一語性表現を自立語とみなしたものであり、概略以下のように表される。

＜接頭語＞*＜自立語＋一語性表現＞＜接尾語＞*＜付属語＋付属語的表現＞*

見掛け文節は拡張文節に対して、漢字や語のゆれによる表記の違う候補を素早く求めるためのもので、概略以下のように定める。

＜自立語＋数詞＋形式名詞＋補助用言＋拡張表現＋接頭語＋接尾語＋冠数詞＋助数詞＞＜助詞＋助動詞＋複合助詞＞*

3．2　派生表記

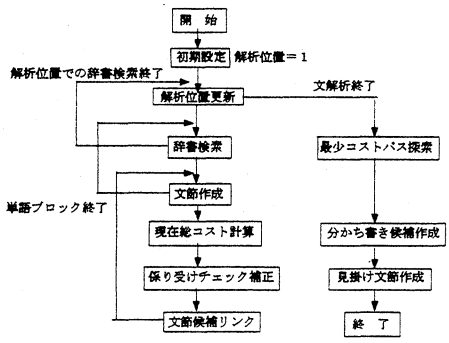
固定的共起表現データの見出し単位に、表記の「ゆれ」情報を総て辞書入れると、データ量の増大と処理速度の低下を招く恐れがある。そこで代表的な表記（代表表記という）とゆれに伴う表記（派生表記という）を、単語の読み・品詞・意味単位にブロック化して持ち、文節分かち書きは代表表記でのみで行い、分かち書き終了後採択された見掛け文節に対しその各候補の派生表記を発生させる。なお、変換後派生表記がユーザにより選ばれ

た場合は代表表記と入れ替えるという学習を行う。派生表記には、次のようなものがある。

送り仮名	売り上げ／売上げ／売上
長音の有無	メモリー／メモリ
拗音のゆれ	バッファ／バッファ
ヴとバ	ヴァイオリン／バイオリン
記号	電話／℡
かな	ください／下さい
カナ	ウナギ／鰻
漢字	浜／濱

3．3　文節分かち書き

文節分かち書き処理は、基本的に文節を単位とするコスト最小法[5]を用いている。これは、文節のコスト、文節間コスト、係り受けコストを計算し、当該文節までの最小のコスト和（最小累積コストと呼ぶ）の最も小さいパスを後方より求め、文節分かち書きを行う方法である。なお、拡張文節により分かち書きが行われ、その後に見掛け文節に分けられる。コスト最小法による文節分ち書きは概略次のような手順で行われる。



（1）コスト計算

拡張文節に対し構成要素によって重み付けを行ったものを文節コストと呼ぶ。文節コストは基準値を２とし、接頭語・接尾語・付属語的表現（複合助詞を除く）のいずれかが存在した場合、１語につき１を加算し最大値を４とする。

修飾関係や漢字の構成要素により文節の連接関係に重み付けを行ったものを文節間コストと呼ぶ。連接する文節コストの和（最小累積コスト）に対し、以下のような場合文節間コスト－１を加える。

連体詞、用言連体形に体言が後接する場合
副詞、用言連用形に用言が後接する場合
名詞と名詞が連接する場合２文字＋２文字
２文字＋１文字
１文字＋２文字

文節間の係り受けをチェックし、係り受けの関係が成立したらその強度によって重み付けを行う。これを係り受けコストと呼ぶ。係る文節と受ける文節の両者の最小累積コストに対して以下のような係り受けコストを加える。

強	固定的共起表現（他語性表現）	－３
↑	緩やかな自立語間の共起	－２
↓	意味と自立語の共起	－１
弱	意味と意味との共起	－１

（2）係り受け補正

最小コスト法では、文節データはその末尾の桁位置に文節長の長い順、最小累積コストの小さい順にリンクされており、文節間の係り受けが成立したら係り語と受け語の両者の最小累積コストが補正され、同一文節末尾で同一文節長のデータの中で、リンク順番も補正される。これにより当該文節候補の優先度が高められる。

（３）見掛け文節の作成

拡張文節による文節分かち書きと見掛け文節による文節分かち書きとその文節候補の例を下記に示す。

例１　人は/気が利くにこしたことはありません
　→　人は/気が/利くに/こした/
　　ことは/ありません

気が/利く：一語性表現
に/こした/ことは/ありません:付属語的表現

例２　第１回/大会に対して/一石を投じる価値がある
　→　第/１/回/大会に/対して/一石を/
　　投じる/価値が/ある

第１回：冠数詞（第）、数詞（1）、
　　助数詞（回）
に/対して：付属語的表現
一石を/投じる：一語性表記
価値が/ある：付属語的表現

なお、拡張文節で「この/気が利く」と分かち書きされるような場合、見掛け文節の採用により「この/気が/利く」としても扱われるため、ユーザは見掛け文節に応じて「木が」、「効く」を次候補選択して。「この/木が/（薬効として）効く」という変換も比較的手早く行うことができる。

５．おわりに

本論文においては、固定的共起表現を用いたかな漢字変換の概要を述べた。拡張文節と見掛け文節という二様の概念を用い、さらに表記のゆれに対処することで操作性の向上を目指しているが、さらに固定的共起表現の拡充と内容の整備が必要である。今後、大量の実験による評価を行っていきたい。固定的共起表現は、かな漢字変換だけでなく付属語的表現の中の助詞的表現による幅広い格の認識や、一語性・多語性表記に意味的に一致する語の対比・書き換えなどにより、機械翻訳などの自然言語処理システムに幅広く活用できる。今後の課題としては、「ほとんど－ない」などの付属語的表現に呼応する自立語表現の収集も必要である。また、固定的共起表現のコーパスからの自動収集と現在収集されたものとの対比・分析も行っていきたいと考えている。

参考文献

[1]首藤,楢原,吉田:日本語の機械処理のための文節構造モデル,電子通信学会論文誌,vol.62-D,NO.12,1979
[2]首藤,吉村,武内,津田:日本語の慣用表現について,情報処理学会研究報告,88-NL-66,1988
[3]首藤,吉村:日本語における語の固定的共起,電子情報通信学会　文法的知識と意味的知識の蓄積、管理シンポジウム論文集,1989-1
[4]首藤,楢原:日本語の文構造のわく組みを与える表現,福岡大学総合研究所報,第６３号抜刷,昭和５８年３月
[5]吉村,武内,津田,首藤:コスト最小法を用いた日本語文の形態素解析,情報処理学会自然言語研究会資料 60-1（1987）
[6]吉村,武内,津田,首藤:未登録語を含む日本語文閏の形態素解析,情報処理学会論文誌 VOL.30,NO.3,1989