辞書見出し語の5文字漢字熟語を対象とした語基構成の解析

郭 恩東，森本 貴之，後藤 智範
神奈川大学理学部情報科学科

1. はじめに

日本語のテキストにおいて、主要な概念・テーマは漢字熟語または漢字熟語を含む名詞句に表現される。特に数文字以上の漢字熟語は、複雑な構造を有し、その構造を分析することは高精度の形態素解析、関連語の選定、未知語の推定などに有効と考えられる。

漢字熟語を対象とした複合語の構造解析の研究は、野村の研究[1][2]を嚆矢とし、大規模な医学用語データベースを対象とした調査[3][4]、記念には意味論的分析を試行した、語彙概念構造に基づく解析の研究[5][6]がある。

本研究は、高速かつ高精度の形態素解析(複合語解析も含む)、あるいは、漢字熟語の構成要素から概念の上位下位に位置する用語、関連語の推定を意図し、一般辞書および専門用語辞書の漢字熟語だけからなる見出し語を対象に、長さ毎に漢字熟語を分類(5 文字～10 文字)し、下記の観点からの分析している。

(1) 構成語基の品詞並び
(2) 構成語基の係り受け解析

本研究では、5 文字漢字熟語を対象とし上記の 2 項目の解析結果について報告する。

2. コーパス

本研究では、表2.1に挙げる辞書の見出し語のうち字種として漢字だけからなる見出し語で5文字の漢字熟語(約2万)から、一般名詞のみを解析対象とした。また、一部として固有名詞を含む用語、仏教用語、故事成語、化学物質名は分析対象から除外した。

表2.1　辞書別漢字熟語数

辞書名	見出し語漢字熟語
広辞苑	137,514
角川類義語辞典	36107
電気・電子情報用語辞典	35,151
EB科学技術用語大辞典	43,8910
コンピュータ用語辞典	29,348語

3. 漢字熟語解析手順

図3.1に語基分割手順を示す。

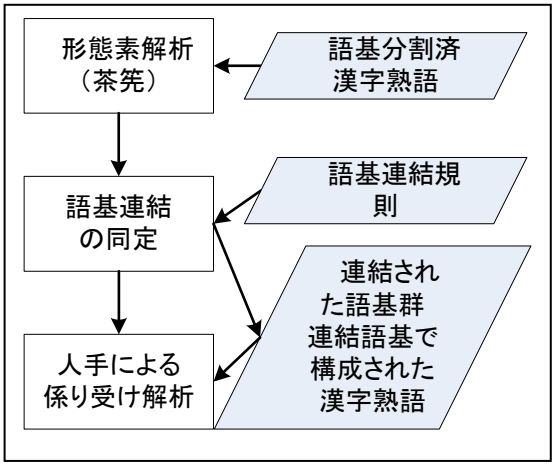


図3.1 漢字熟語の語基分割手順

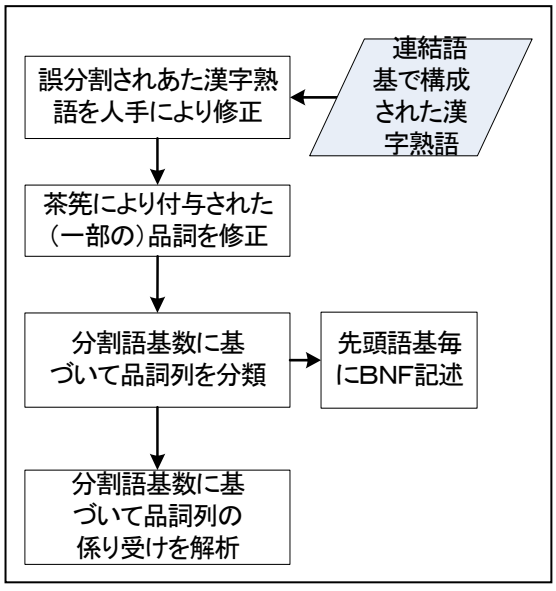


図3.2 語基構成解析手順

茶筅の品詞カテゴリー体系は階層が深く、品詞並びが発散する可能性があったため、本解析では、学校文法の近い下記の品詞を用い、一部は人手により修正した。

名詞　動詞　サ変　形容動詞語幹
形容詞語幹　接頭辞　接尾辞　副詞
動詞　接続辞

4．結果

対象間熟語は 2、3、4 語基のいずれかに分割された。

4.1 品詞列パターン

表 4.1 に構成語基数毎に出現した品詞列パターンの種類を挙げる。先頭語基毎に出現した品詞列パターンに対して BNF 標記および、相当する漢字熟語の例をそれぞれ挙げる[7]。

表 4.1　構成語基数毎の品詞列パターン数

構成基数	2	3	4
品詞列パターン数	6	57	51
比率(%)	5	50	45

4.1.1　2 語基

構成要素が 2 語基となった漢字熟語の品詞列パターンは下記の 6 種類であった。先頭語基毎の BNF 表記を下記に記載する。

<サ変>(<サ変>|<名詞>)
<形容動詞語幹><名詞>
<名詞>(<サ変>|<形容動詞語幹>|<名詞>)

4.1.2　3 語基

先頭語基の品詞、当該品詞で始まる品詞列の種類およびその比率を表 4.2 に挙げる。

表 4.2　3 語基の先頭品詞毎の品詞列パターンの種類

	サ変	形動語幹	接頭辞	動詞	副詞	名詞
種類	17	9	8	4	1	18
比率	30	16	14	7	2	31

使用された品詞のうち、先頭語基に現れた品詞は6種類であった。品詞列パターン全体の60%以上が名詞とサ変名詞で始まることがわかる。

下記に、動詞で始まる語基の品詞列パターンのBNF表記とその実例を挙げる。

<先頭語基が動詞の3語基の5文字熟語> ::=
<動詞>{　[<サ変> (<サ変> | <名詞>)] | [<接頭> (<名詞>)]

[実例]　「誤動作警報」　→　誤 動作 警報

4.1.3　4 語基

3 語基と同様のデータを表 4.2 に挙げる。

表 4.2　3 語基の先頭品詞毎の品詞列パターン

	サ変	形容	形動	接頭辞	数詞	動詞	名詞
種類	1	1	1	15	19	1	13
比率	2	2	2	29	37	2	26

先頭語基の品詞は下 7 種類であった。3 語基と比較すると、形容詞語幹および数詞が加わり、副詞で始まる語基が無いことがわかる。下記に、形容詞語幹で始まる品詞列パターンの BNF 表記とその実例を挙げる。

<先頭語基が形容詞語幹の4語基の5文字熟語> ::=
<形容詞語幹>[(<接頭辞><サ変><接尾辞>})]

[実例]　「完全無肢症」　→　完全 無 肢 症

4.2　係り受けパターン

次節以降に、構成語基数毎の係り受けパターンと品詞列パターンとの出現傾向についての結果を示す。

4.2.1　2 語基構成熟語

2 語基については､当然ではあるが全て図 4.1 の係り受けパターンと一致した。

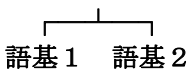


図 4.1　2 語基の係り受けパターン

4.2.3　3 語基構成熟語

表 4.1 に挙げた､57 種類の品詞列パターンについて、人手による係り受け構造を調査した結

果、図 4.2 に示す 2 種類の係り受けパターンが見られた。

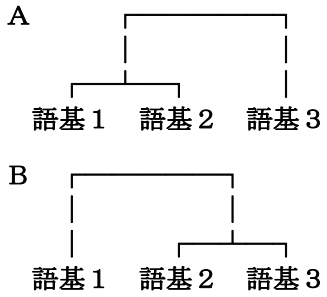


図 4.2　3 語基の係り受けパターン

表 4.1 は先頭語基の品詞によってどの係り受けパターンとなるかを示している。

表 4.1　3 語基構成熟語の係り受け構造
　　毎の先頭語基品詞による品詞列数

係り受け	A	B
サ変	11	6
形動	4	5
接頭辞	8	0
動詞	3	1
名詞	3	15
副詞	1	0

4.2.4　4 語基構成熟語

同様に 4 語基構成熟語の 51 種類の品詞列パターンについて調査した結果、図 4.3 に示す 4 種類の係り受けパターンが見られた。

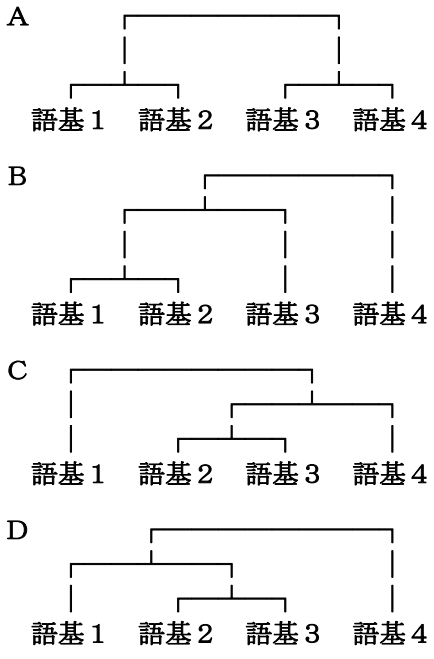


図 4.3　4 語基の係り受けパターン

表 4.3　4 語基構成熟語の係り受けパターン
　　毎の先頭語基品詞による品詞列数

係り受け	A	B	C	D
数詞	11	7	0	3
サ変	0	0	1	0
形容	0	0	0	1
形動	0	0	0	2
接頭辞	8	4	0	4
動詞	0	0	0	1
名詞	3	3	2	6

図 4.3 に示す 3 種類の係り受け構造は、今回のコーパス中には出現しなかった。

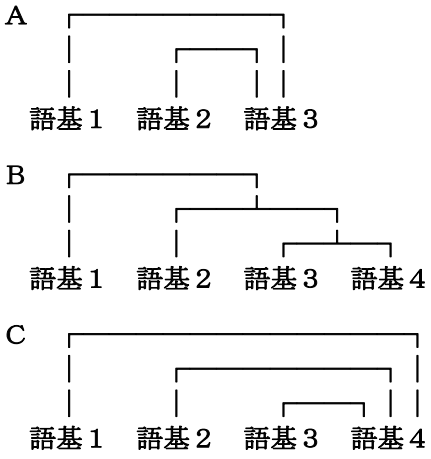
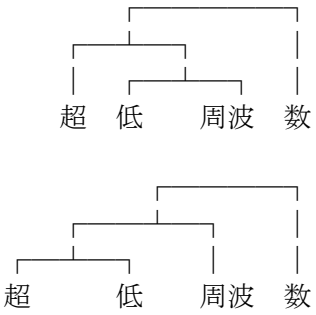


図 4.3　出現しなかった係り受けパターン

5. 考察

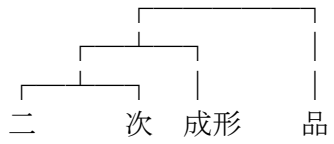
人間による係り受け決定の段階で、幾つかの漢字熟語は1つの構造が決定困難な事例があつた。下記にその実例を挙げる。

実例：　超 低 周波 数
係り受け：

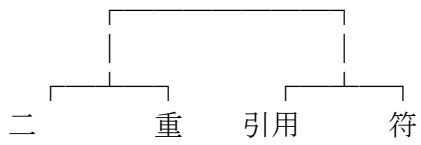


また、表4.3において、先頭語基の品詞毎の係り受けパターンは、表4.2のそれとは一致しない。これは、1種類の品詞列並びが、漢字熟語そのものにより、複数の係り受け構造をとる実例が存在するためである。下記のそれぞれ2つの漢字熟語の品詞列並びは同一であるが、それぞれの係り受け構造は異なっている。

品詞列：　数詞 接尾辞 サ変 名詞
実例：　　二　次　成形　品



品詞列：数詞 接尾辞 サ変 名詞
実例：　二　重　引用　符



このような事例は、3語基構成の漢字熟語では見られなかった。

5. 終わりに

4章の結果から明らかなように、品詞列パターンが決定すると、係り受け構造は非常に高い確率で決定できる。宮崎による自動複合語分割手法では、全ての分割語基リストに対して、係り受け解析が必要とされる[8]。一方、本研究結果を拡張し、長さ毎の漢字熟語についての品詞列パターンのデータが網羅的に得られれば、係り受け解析を行わずに、先頭の語基から順に語基単位、品詞が認定されるにつれに、制約がせばまり、高精度かつ高速で、自動分割およびその構文構造の決定がが可能となると想定される。

註・参考文献

[1] 野村雅昭. 三字漢字の構造. 秀英出版. 国立国語研究所報告. No.51. pp.37-62(1973).

[2] 野村雅昭. 四字漢字の構造. 秀英出版. 国立国語研究所報告. No.54. pp.36-80(1974).

[3] 小山照夫, 大江和彦. 医学専門用語の構造解析. 学術情報センター紀要. No.6, pp.115-124(1994).

[4]小山照夫, 大江和彦. 日本語医学専門用語の構造解析. 情報知識学会第 2 回研究報告会講演論文集. pp.17-20(1994).

[5] 竹内孔一, 内山清子, 吉岡真治, 影浦峡, 小山照夫. 語彙の制約を考慮した複合語解析モデルの構築. 情報処理学会. 情報学基礎研究会報告. No.57, pp.71-78(2000).

[6] 竹内孔一, 内山清子, 吉岡真治, 影浦峡, 小山照夫. 語彙概念構造を利用した複合名詞内の係り関係の解析. 情報処理学会論文誌. Vol.43, No.5, pp.1446-1456(2002).

[7] 紙面の制約により全てのBNFを列挙することはできない。同様の理由により、比較的単純な表記となるBNFを記載した。

[8] 宮崎正弘. 係り受け解析を用いた複合語の自動分割法. 情報処理学会論文誌. vol.25, No.6, pp.970-979(1984).