E_038

日本語文末表現の取り扱いについて
On the Processing of Japanese Sentence-final Expressions

田辺　利文*　本田　聖晃*　高橋　雅仁**　小山　泰男***
Toshifumi TANABE　Takaaki HONDA　Masahito TAKAHASHI　Yasuo KOYAMA

吉村　賢治*　首藤　公昭*
Kenji YOSHIMURA　Kosho SHUDO

1　はじめに

自然語文の非命題的内容(Non-Propositional Content; NPC)とは、相、態、様相などの時間的、偶発的、認識的、発話行為的情報を自然語文の命題的内容(Propositional Content; PC)に添え、発話を現実化するものと考えられている。日本語では、NPC は文末の助動詞、終助詞、その他の複単語表現(Multi-Word Expression; MWE)で表される場合が多い。我々はそれらの表現を非命題的内容指示語(NPC Indicator; NPCI)と呼び、網羅性を重視した日本語 NPCI 辞書を構築してきた。本論文では、これらの表現で、特に MWE の NPCI を必要に応じて単語として取り扱うことにより、日本語述部の構文構造とその非命題的意味構造(NPS)との明瞭な対応がとれること、および NPS を入れ子構造と捉えることで、意味・構文的に幅広い文末表現が取り扱える事を示す。さらに、NPCI がどの程度日本語文に現れるかの統計データを示し、NPCI の正しい取り扱いが重要であることを示す。

2　非命題的内容指示語 (NPCI)

2.1　NPCI 辞書の概要

日本語文における非命題的内容(NPC)は、多くの場合、文の末尾の助動詞、終助詞その他多種の複単語表現(MWE)により表される。これまで我々は約 1,450 の複単語 NPCI を収集したが、それらの表現は、次に示す 3 つのうち少なくとも 1 つの性質を持つ単語列である。

f_1: 熟語性、f_2: 語彙的一体性、f_3: 確率的束縛性

熟語性とは、構成性原理の成り立ちにくさを意味しており、構成している単語の通常の意味から全体の意味を構成するのが難しいことを意味する。語彙的一体性とは、分離しにくさ（要素単語の間への他単語の割り込みにくさ）を、確率的束縛性とは、要素単語相互の確率的な縛りの強さを意味する。各複単語NPCIについて、これらの性質を有無を 3 つ組<f_1 f_2 f_3>によって記す。但し、f_iは現在 2 値としている。例えば、熟語的であり、分離可能で、確率的に結びついていないNPCI “必要·が·ある” には<100>、また、熟語的であり、分離不可能で、確率的に結びついているNPCI “なければ·なら·ない” には<111> を与え、複単語表現としての基本的な性格を記述する。ここで、記号 ‘·’ は通常の単語境界を表す。我々のNPCI辞書は約 1,450 個の複単語表現と 50 個の助動詞、終助詞を収録している[1]。筆者らは網羅性の高さを目標の一つと考えており、例えば、(森田ら, 1997)の収録表現のうち該当表現は本辞書ですべてカバーされている。

2.2　NPCI の熟語性

次の文の意味を考える。

(1) “彼·は·そこ·に·居る·べきで·なかっ·た”

動詞“居る”に後接した助動詞“べきだ”、 “ない”、“た”のそれぞれの意味は OBLIGATION、NEGATION、PAST-TENSE と考えられる。 そこで、これらの意味を記号化して、この文の非命題的意味構造を

(2)PAST-TENSE [NEGATION$_1$ [OBLIGATION$_2$[“彼·が·そこ·に·居る”]]]

と表してみよう。非命題的意味をこのような単純な入れ子の構造で捉える事には機械処理を単純化する点で大きな魅力がある。個々の意味の記号を非命題的意味関数(NPF)と呼ぶ。

次に(3)のような例を考える。

(3) “彼·は·そこ·に·居る·べきだ·と·いう·こと·は· なかっ·た”

(3)も骨格となる文（命題的文）“彼がそこに居る” に OBLIGATION の “べきだ”、NEGATION の “ということはない”、PAST-TENSE の “た” が同順に現れているのでその意味は(2)で表せる。しかし、明らかに(1)は(3)と同義ではなく、むしろ次の(4)に近いと考えられる。

(4) “彼·が·そこ·に·居·た·の·は·まずい”

(4)の意味は(2)の記法に従えば(5)のようなものになろう。

(5)NEG-EVAL[PAST-TENSE[“彼·が·そこ·に·居る”]].[2]

そこで、(1)の意味を(2)でなく(5)として導くにはどうするかが問題となるが、(1)の単位切りから構成的に(5)を求める事には相当な困難が予想される。

* 福岡大学
** 久留米工業大学
*** （株）セイコーエプソン

[1] “すら”、“こそ”、“さえ” などのような、付加的情報を与える非命題的後置詞は、NPCIの範疇に入れていない。このような非命題的後置詞は一般的には取り立て助詞といわれる。

[2] NPF間の微妙な違いや程度などを区別するためNPFの右下に数字を入れる。

n	述部の生起	単一の単語 NPCI の生起 (A)	複単語 NPCI の生起(B)	A+B	B/(A+B)
0	4,899	-	-	-	-
1	3,131	1,852	1,279	3,131	0.408
2	966	1,128	804	1,932	0.416
3	178	276	258	534	0.483
4	34	63	73	136	0.537
5	2	5	5	10	0.500
合計	9,210	3,324	2,419	5,743	0.421

表 1 述部 PRED· m_1· m_2· … · m_n 内のNPCIの生起

筆者らは、“べきで·なかっ·た” を結合したNPF、NEG-EVAL[PAST-TENSE[x]][3] を与えるNPCIとして扱うことにした。このように、一つのNPCI[4]によってNPFの結合が一括して与えられることも許す。

2.3 NPCIの統計データ

日本語文でどの程度 NPCI が述部に現れるか、また、現れた NPCI がどの程度複単語表現であるかを EDR コーパス (EDR, 1996)からランダムに抽出した 9,210 個の文を対象にし、述語の品詞が動詞、形容詞の場合に限って調査した。調査の結果を表 1 に示す。n は述語に連続して現れた NPCI の個数である。その結果、文末に少なくとも 1 個の NPCI が含まれる割合は 47%(=(9,210-4,899)/9,210)であった。また、述部に NPCI がある場合、それが複単語表現である割合は 42%であった。述部にあった連続した NPCI の個数は最大で 5 であった。NPCI の個数が 5 であるものの例を以下に示す。下線で示しているものが NPCI であり、そのうち太字で表しているものが複単語の NPCI である。

“左右/さ/れ/ない/**よう·に·し**/たい/**もの·だ**”
“発揮/でき/ない/**状況·に·ある**/**から·だ**/**と·いう**”

3 非命題的意味構造(NPS)

3.1 日本語文の構造の概形

複単語の NPCI を適切に設定することで、日本語文の構造の大枠を次の生成規則で表現することができる。

(6) $S_0 \rightarrow BP^*$· PRED,
(7) $S_i \rightarrow S_{i-1}$·$NPCI_i$, $(1 \leq i \leq n)$,

ここでS_0 、 BP 、 ‘*’ はそれぞれ骨格文、文節、閉包演算子を表す。以後、PRED· $NPCI_1$· $NPCI_2$ · … · $NPCI_n$ を文の述部と呼ぶことにする。

3.2 文の NPS

自然語文の NPS は、次のような入れ子型表現で表すことができる。

(8) $M_m[M_{m-1}…[M_2[M_1[S]]]…]$,

但し、S は命題的な骨格文、M_i $(1 \leq i \leq m)$, はNPFである。我々は、150 個程度のNPF、例えば GUESS, NECESSITY, OBLIGATION, INCHOATIVE, PERMISSIVE, POLITENESS, IMPERATIVE, PROGRESSING, PASSIVEなどを提案している。これらのNPFは当面、日常会話程度の日英機械翻訳を目的としたものである。我々は、NPFを、時制、相、態、様相、発語内行為など 9 つのカテゴリに分類している (Shudo et al., 2004)。
NPS には言語依存性がないと考えられることから、NPS は言い換えや機械翻訳を行う際の中間表現として有効であると考えている。
また、それぞれの NPCI は、何らかの対応する NPF を持つように決められており、(6),(7)に示される構文構造と、(8)に示される意味構造とには一種の同型性があるといえる。NPCI が文の述部にいくつも並んだ複雑な文末表現の場合でも、NPF との対応をつけることにより、NPS を求めることが可能であるといえる。このように(8)は構造のシンプルさと同時に対応可能な表現の多様さの点で工学的に重要な性質を示していると考えている。

4 実験

どの程度正しく NPS を構築できるかを求めるために、独自に開発した複単語表現を組み込んだ分かち書きシステム、およびNPS構築システムを用いて解析実験を行った。
EDR コーパス(EDR, 1996)からランダムに抽出した、動詞を主述語とする述語文節 1,481 個に対して、分かち書きシステムによる正しい出力は 1,356 個であり、誤った出力は 125 個であった。誤った出力 125 個のうちの 1 個は chasen でも正しく分かち書きされなかった。複単語表現を組み込んだ我々の分かち書きシステムは、最小コスト法に基づくアルゴリズムを採用している。そのため、本来、複単語表現ではない表現を複単語表現とみなしてしまう、複単語表現の過認識の問題がある。

[3] 他の手段としては、NPFがPROHIBITION$_2$であるような、例より短い複単語表現“べきで·ない”を適用し、次にPAST-TENSEである “た” を適用してPAST-TENSE[PROHIBITION$_2$[“彼·が·そこ·に·居る”]]とし、最終的に PAST-TENSE[PROHIBITION$_2$[x]をNEG-EVAL[PAST-TENSE[x]]とする方法もあるが、過剰生成に注意する必要がある。
[4] 他の典型的な例が“まい”であり、これは一般的な助動詞であるが 2 つのNPFの合成 GUESS$_2$[NEGATION$_1$[x]].で表される。

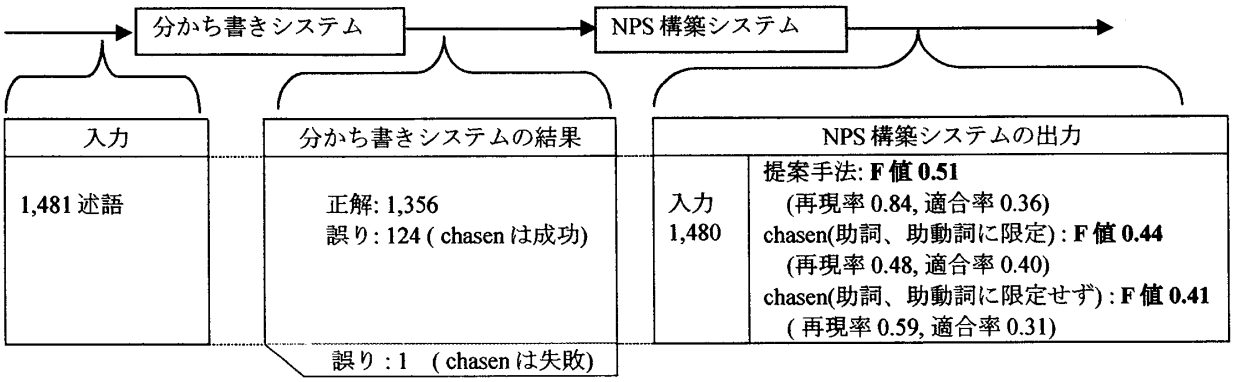

図 1: 全体の述語文節(1,480 個)に対する評価の概略

過認識の例として、“それは今日私が買ったものだ”中の“たものだ”を単一の NPCI と認識してしまう。とりあえず本実験では、このような複単語表現と過認識した出力も、正しい分かち書き出力の中に含めた。
4.1 では正しく分かち書きされた述語文節 1,356 個のみを扱った場合の実験結果を示し、4.2 にて、述語文節全体の結果を示すことにする。

4.1 正しく分かち書きされた述語文節のみの実験結果

正しく分かち書き出力された 1,356 個の述語文節に対して、NPS 構築システムを用いて NPS を求めた。その結果、構築された NPS は 3,149 個で、そのうち正しい NPS は 1,247 個であり、再現率 0.92、適合率 0.40、F 値 0.55 であった。
複単語表現を組み込んだ手法との性能比較のため、正しく出力された 1,356 個の述語文節に対し、chasen(chasen, 1996) の分かち書き結果を用いて NPS を構築した。NPCI の品詞は助詞、助動詞に相当するため、まず、chasen が出力した述語文節内の単語が助詞、助動詞の場合に限定して NPF を割り当てることで NPS を構築した。その結果、構築された NPS は 1,647 個で、そのうち正しい NPS は 591 個で、再現率 0.44、適合率 0.36、F 値 0.39 であった。また、NPCI の中には、chasen が助詞や助動詞以外の出力をする単語も含まれ得ることを考慮し、述語文節内の主述語以外の単語全て、つまり助詞、助動詞に限定しないで単語に NPF を割り当てて同様に行ったところ、構築された NPS は 2,585 個で、そのうち正しい NPS は 748 個で、再現率 0.55、適合率 0.29、F 値 0.38 であった。

4.2 全体の評価

4.1 での結果は、正しく分かち書きされた述語文節 1,356 個のみの結果であった。しかし、システム全体の評価としては、1,356 個の結果だけでなく、分かち書き結果が正しくなかった述語文節 125 個により構築される NPS についても考慮する必要がある。そのため 1,481 個の述語文節から、chasen でも分かち書きに失敗するとみられる 1 個を除外した 1,480 個の述語文節に対する全体の評価を、chasen を用いた場合と複単語表現を組み込んだ手法とで比較した。ここでは複単語表現を組み込んだ分かち書きシステムで失敗した 124 個の述語文節に対して NPS を仮想的に構築することで全体の評価とした。
chasen を用いた仮想的 NPS 構築に際しては、解析実験の結果を一部用いた。述語文節内の単語が助詞、助動詞のみに限定して NPF を割り当てる場合には、124 個の述語文節から 151 個 (=124*1,647/1,356) の NPS が構築され、かつ、それぞれの述語文節から生成される NPS には必ず 1 つの正解が含まれるとすると、1,480 個の述語文節に対する評価値として再現率 0.48、適合率 0.40、F 値 0.44 が得られた。同様に、助詞、助動詞に限定しない場合では、再現率 0.59、適合率 0.31、F 値 0.41 が得られた。
一方、複単語表現を組み込んだ手法に対しては、124 個の述語文節から 288 個(=124*3,149/1,356)の NPS が構築されるものとし、かつ、ここで構築される NPS はいずれも不正解であるから、1,480 個の述語文節に対する評価値として再現率 0.84、適合率 0.36、F 値 0.51 が得られた。これらの結果により、複単語表現を組み込んだ分かち書きシステムの出力を用いて NPS を構築すれば、chasen を用いた場合と比較して質の良い NPS を得られることが全体の評価から示された。また、再現率は全体の評価でも 0.84 であったことから、NPCI および NPF の網羅性も比較的高いものと思われる。
図 1 に全体の述語文節(1,480 個)に対する評価の概略を示す。

5 関連研究

日本語文末表現とその意味に関する関連研究をいくつか紹介する。(横野, 2005)は、日本語の複単語文末表現が話者の感情の推定に役立つことを少数の NPCI を用いた実験で示した。(Shirai et al., 1993) は、日英翻訳システムの前処理としての言い換え処理のために NPCI を取り扱っている。しかし扱った NPCI の網羅性と NPCI の意味は論文中には明記されていない。(益岡ら, 1989) は、日本語の NPCI を 3 種に分類し、階層的な関係があるとして、次のような構造を提案している。

(9) $[[[S]E_3]E_2]E_1$

ここで、Sは命題、E_1は真偽判断に関する表現、E_2はテンスに関する表現、E_3はみとめ方に関する表現[5]である。但し、(9)の構造は単に表現 3 種間の階層関係を記述しているだけでありNPSのような再帰的構造を記述しているわけではなく、また表現の詳細には触れられていない。

[5] 否定の「ない」などがある。

(村田ら, 2005) では、日本語文において、1文に対し、文末からみた1文字目から10文字目までの10種類の文字列と文が持つ時制、相、様相情報を SVM を用いて学習させた結果、時制、相、様相情報を高い精度で推定したことを報告している。しかし、村田らの手法は、i) 想定した NPF の種類が34種類と少ないため文末表現が持ち得る意味の網羅性に欠けること、ii) 10文字以上からなる述部は扱えないこと、などがある。特に ii)は、日常会話文では文末に NPCI が多く連接して用いられることが多いため、入れ子型の NPS を採用した本提案手法が意味のもれをカバーできると思われる。

6 おわりに

自然語文の非命題的内容は、対話理解、文脈モデル、話者の態度の推定などの自然言語処理で重要な役割を果たす。本論文では日本語文の非命題的内容を取り扱うための枠組みを紹介した。この枠組みは、熟語性の複単語表現を一つの単語として取り扱うことで、日本語の文末の構文的構造とその非命題的意味構造とを対応付けることができることを示した。

また、日本語の述部において調査した結果、複単語表現としての NPCI は、NPCI 全体の 42%生起していることも明らかにした。(Sag et al., 2002) に述べられている、WordNet 1.7 (Fellbaum, 1999) における英語見出しの複単語表現の生起の割合が 41%であったことと併せると類似性が興味深い。また、本モデルに基づく実験システムを使って、NPCI がどれだけシステムに認識されるかの簡単な調査を行い、かつ、入れ子型の NPS の枠組みの必要性を述べた。

今後の課題としては、NPCI と NPF の網羅性の向上や、同じ NPF をもつ NPCI 同士の言い換え、NPF 列の相互変換、同じ NPF を持つ NPCI が複数ある場合どの NPCI を標準とするか、対話処理やメール分類などの応用処理を行うことなどが考えられる。また、第 4 章でも述べたように、複単語表現である NPCI には、複単語表現の過認識の問題もある。例えば、“私はそこでよく花を買ったものだ”では、CUSTOM という NPF をもつ NPCI “たものだ”を認識すべきであるが、“それは今日私が買ったものだ”では“たものだ”を NPCI と認識すれば誤りとなる。どのような場合に複単語表現として認識するべきかの情報の整理など、解決すべき課題は多く残されている。

参考文献

chasen(茶筌). 1996. 形態素解析システム茶筌. http://chasen.naist.jp/hiki/ChaSen/

EDR(日本電子化辞書研究所). 1996. EDR 電子化辞書 . http://www.iijnet.or.jp/edr/

Fellbaum, Christine, ed.: 1998. *WordNet. An Electronic Lexical Database*. Cambridge, MA: MIT Press.

横野. 2005. 情緒推定のための発話文の文末表現の分類. 情報処理学会研究報告, NL-170-1:1-6.

Iwan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. *Multiword Expressions: A Pain in the Neck for NLP*. The Proc. of the 3rd CICLING: 1-15.

Kosho Shudo, Toshifumi Tanabe, Masahito Takahashi and Kenji Yoshimura. 2004. *MWEs as Non-propositional Content Indicators*. The Proc. of the ACL2004 Workshop on Multiword Expressions: Integrating Processing: 32-39.

森田良行, 松木正恵. 1989. 日本語表現文型用例中心・複合辞の意味と用法, アルク.

村田真樹, 内元清貴, 馬青, 金丸敏幸, 井佐原均. 2005. テンス・アスペクト・モダリティの翻訳における機械翻訳システムの誤りの調査. FIT2005: 77-80.

益岡隆志. 1989. モダリティの構造と疑問・否定のスコープ. 日本語のモダリティ. 仁田義雄,益岡隆志(編). くろしお出版: 193-210.

Satoshi Shirai, Satoru Ikehara and Tsukasa Kawaoka. 1993. *Effects of Automatic Rewriting of Source Language within a Japanese to English MT System*. Fifth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-93: 226-239.