日本語処理のための談話指標・文副詞的複単語表現レキシコン：JMWEL_DM/SA

首藤公昭[†1]　田辺利文[†2]　髙橋雅仁[†3]

概要：「それはそうと」，「何故なら」，「要するに」，「分かり易く言えば」，「驚くべきことに」，「話を戻すと」など，日本語文の文頭で独立して用いられ，話の流れを制御したり後続文の評価を与えたりして文書の読解や対話を円滑にする定型表現，複単語表現（Multiword Expression, MWE）約 1,400 を収録したレキシコン JMWEL_DM/SA について報告する．

キーワード：談話指標，語用指標，文副詞，機械辞書，言語資源，日本語処理，機械翻訳

A lexicon of discourse marking or sentence adverbial multiword expressions for Japanese language processing: JMWEL_DM/SA

KOSHO SHUDO[†1]　TOSHIFUMI TANABE[†2]
MASAHITO TAKAHASHI[†3]

Abstract: We introduce, in this paper, a new lexicon of Japanese discourse marking or sentence adverbial multiword expressions: JMWEL_DM/SA. The lexicon, which is tuned for discourse processing, is a sub lexicon of the large scale lexicon of Japanese Multiword Expressions: JMWEL, which has been developed for the linguistically precise, wide-coverage and phrase-based Japanese language processing.

Keywords: Discourse marker, Discourse relation marker, Pragmatic marker, Sentence adverb, Discourse understanding, Dialogue system , Multiword Expression, Lexicon for NLP, Linguistic Resource for NLP

1. はじめに

「それはそうと」，「正直に言って」，「何故かというと」，「そうはいっても」など，文頭で用いられ，先行する文（あるいは段落）と後続の文（あるいは段落）との意味的関係を表示したり，後続文の記述の情報・評価を先行提示したりして，文書の読解や対話を円滑にする表現を談話指標（discourse marker）・文副詞（sentence adverb）的表現と呼ぶ．従来，この種の表現として「なお」，「だが」，「しかし」，「だから」，「さて」のような単語の接続詞，副詞，間投詞が考察されてきたが，近年，上記のような長単位の定型表現にも目が向けられている([3][4]等)．しかしながら，この種の表現の全体像を提示する辞書化は，筆者らの知る限り，まだ行われていない様である．

計算言語学の世界では，今世紀に入り，自然言語に頻出する長単位定型表現を一般的に複単語表現（multiword expression: MWE）と名付け[5]，機械処理におけるその重要性が認識されるに至っており，例えば，計算言語学会 ACL を中心に毎年 MWE workshop が開催されている[8]．

言語表現の定型性については，言語学の世界でも今世紀に入り，定型言語（formulaic language）[2]，語彙連鎖（lexical bundles）[1]等の枠組みで種々の研究が進められている．

本項で解説する JMWEL_DM/SA は，大規模日本語複単語表現レキシコン(JMWEL)[7]から談話指標・文副詞的と思われる約 1,400 表現を抽出し，見出し，分かち書き情報，異表記情報，構文機能情報等，構文構造情報，文脈条件等を与えた計算機処理用レキシコンである．

2. 表現の採録

表現の採録基準は，

1. 意味の一体性(semantic unity)
2. 要素語間の強い共起性（確率的親和性 probabilistic affinity，決まり文句性 collocationality）
3. 構文，意味の非構成性（non-compositionality，イディオム性 idiomaticity）

であり，疑わしいものはなるべく採録するという再現性を重視した編集を目指している．

表現は，不特定多数の日刊紙，雑誌，小説，テレビ，ラジオ報道文などの生データ，および，不特定多数の辞書，事典類から編者の内省によって取捨選択・採集されている．

本レキシコンの見出しは原則として，丁寧語，方言，古語を除く書き言葉現代日本語とする．

†1 福岡大学名誉教授
Emeritus, Fukuoka University
†2 福岡大学
Fukuoka University
†3 久留米工業大学
Kurume Institute of Technology

3. 本レキシコンの特徴

本レキシコンの特徴は,
1. 漢字, カタカナ用法や送り仮名の有無など, 表記の多様性に配慮している
2. 表現内部の文節内連結構造, 係り受け構造, および, 並列構造を記載し, 通常の構文処理とのリンクに配慮している
3. 表現の柔軟性を確保するため内部修飾可能性(internal modifiability) を表現ごと個別的に記載している
4. 構文的に不完全な表現(ill-formed expressions)にも空(省略)要素記号を補って構造記述を与えている
などである.

4. 記載情報

本レキシコンは, Microsoft Excel で作成した xlsx ファイルに纏められており, 表の 1 行に割り当てた 1 個の見出し表現に対して, A～I 欄に以下の情報を記載している. 例えば, 表現「何故かと言うと」に与えた情報は,
A 欄 --- DM/SA
B 欄 --- なぜかというと
C 欄 --- なぜ-か-と-いう-と
D 欄 --- 何故-か-と-(言/云/謂)う-と
E 欄 --- DM/SA_Vto
F 欄 --- [[[Dka]to]V30]to
G 欄 --- <top-of-sentence>*
H 欄 --- <punct. concat.>
I 欄 --- ナシ
以下, 各欄の内容を解説する.

4.1 種別(A 欄)

表現が談話指標・文副詞的である事を「DM/SA」と記し, 他レキシコンとの統合利用の際の標識とする.

4.2 見出し(B 欄)

平仮名ベタ書き見出しを与える. 漢字が複数の読みを持っている場合は可能な読み毎に見出しを与えている. 例えば,「よせばよいのに」と「よせばいいのに」の両方が見出しとされている.

4.3 分かち書き(C 欄)

B 欄の表現がどのような形態素, すなわち, 単語, 接辞(接頭語, 接尾語, 接頭造語要素, 接尾造語要素)から構成されているかを形態素分かち書きで与える. 造語要素とは造語能力が比較的強く, 単独で用いられることのない形態素で, 音読みの一漢字である.
接辞を使った収録表現には, 例えば, 以下が有る.

お : ぼうとう-で-お-はなし-し-た-よう-に
ら : それ-ら-に-より
らく : うらむ-らく-は
ふう : そんな-ふう-に-し-て
がましい : さし-で-がましい-こと-を-いう-よう-だ-が
てき : ぐたい-てき-に-は

形態素の区切りは, ハイフン「-」あるいはアンダースコア「_」で行なっている. 比較的明確な区切りにハイフン「-」, 区切りの可能性のあるところにアンダースコア「_」という 2 段階表示である. また, 明らかに単語であっても, その一部が異なった字種で表記可能であれば, 字種の変わり目にアンダースコアを入れている. 例えば,「くちはばったい様だが」の「くちはばったい」は,「口幅ったい」や「口はばったい」と表記されることがあるため,「くち_はばったい」と弱く区切っている. この情報と D 欄に与えた字種情報「口_幅ったい」とから「口幅ったい」,「口はばったい」,「くち幅ったい」,「くちはばったい」という 4 つの表記が導出できる.

4.4 異表記(D 欄)

片仮名表記, 漢字表記, 送り仮名の有無など, 表記の多様さをコンパクトに記載した欄である. 漢字, カタカナなど, 異表記可能な表現には, C 欄の分かち書きの上で選択肢を与える. 例えば,「何故-か-と-(言/云/謂)う-と」の「(言/云/謂)」の部分は 3 つの漢字「言」,「云」,「謂」のいずれか 1 つが使用可能であること,「断(わ)っ-て-置く-けど」は「断わっ-て-置く-けど」,「断っ-て-置く-けど」の 2 つの可能性があることを表す.

4.5 構文機能と形態的特徴(E 欄)

表現の機能が談話指標・文副詞的であることを DM/SA_α と形式化し, α 部に表現の形態的特徴を与えている. α 部は, 例えば「なぜかというと」の場合, 表現末尾「いうと」が動詞「いう」＋接続助詞「と」であることを Vto と表示している. α の種類は多岐に亘るが, 大別すれば, 下記の通りである.

1. 先行語句が欠落した不完全句: 「だ-けれど」,「と-は-(言/云/謂)う-ものの」
2. 後続語句が欠落した不完全句: 「其れ-は-兎_も_角」
3. 形容詞(的表現), あるいは形容詞(的表現)に付加部が後接したもの:「話せ-ば-長い-けど」
4. 接続詞, あるいは接続詞に付加部が後接したもの: 「しかし_ながら」
5. 副詞, あるいは副詞に付加部が後接したもの: 「其れ-は-然う-だ-けど」
6. 感動詞: 「成る_程」
7. 形容動詞型表現, あるいは形容動詞型表現に付加部が後接したもの: 「具体-的-に-は」
8. 名詞型表現, あるいは名詞型表現に付加部が後接したもの: 「不思議-な-事-に」
9. 動詞型表現連用形: 「(之/是/此れ)-に-対し」
10. 動詞型表現仮定形:「其れ-は-と_も_あれ」
11. 動詞型表現命令形: 「(何/孰)れ-に-せよ」
12. 動詞型表現に付加部が後接したもの: 「何故-か-と-いえ-ば」

4.6　構文構造(F 欄)

C 欄のハイフンによる分かち書きに従って表現内の**係り受け構造**を，修飾子，被修飾子の対をカッコ[ ]で括って記載する．すなわち，句 α の主辞が句 β の主辞を修飾して出来た句 αβ の構造記述を α，β の構造記述 a, b を使って[ab]とする．

ベースとなる構成単語の構造記述は，以下の通りとする．

- 接頭語：　P
- 接尾語：　S
- 接頭造語要素：　Q
- 接尾造語要素：　R
- 名詞：　N
- 動詞：　V（未然形 V11,V12，連用形 V22,V23，終止形 V30，連体形 V40，仮定形 V50，命令形 V60）
- 形容詞：　A （未然形 A13，連用形 A22，A23，終止形 A30，連体形 A40，仮定形 A50，命令形 A60）
- 形容動詞（語幹）：　K00
- 副詞：　D
- 連体詞：　T
- 機能語及び機能性自立語：英小文字
- 接続詞：　C

文節内の語の接続も，便宜上，左 2 分岐句構造とみなして同様の記述を行っている．

例えば，前節のタイプ 8 の表現「斯う-し-た-状況-の-下」の構造記述は，

「斯う」：　副詞 D
「し」：　動詞「する」の連用形 si
「た」：　助動詞「た」の連体形 ta
「状況」：　名詞 N
「の」：　連体格助詞　no
「下」：　名詞　N

を使って[[[[[Dsi]ta]*N]no]N]と記載する．図 1 にその意味する構文木，文節内構造，係り受け構造を示す．

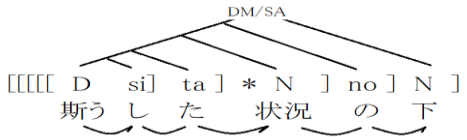

図 1 「斯う-し-た-状況-の-下」の構造

Figure 1　Syntactic structure denoted by [[[[[Dsi]ta]*N]no]N] given to the DM/SA expression「斯う-し-た-状況-の-下」

また，**並列構造**は，括弧表現<　>または《　》，並列される要素は括弧(　)で表わす．例えば，「其れ-や-(是/之/此れ)-や-で」の句表示 [《(N)ya(N)ya》de]は図 2 の構造を意味する．並列要素は「其れ」と「(是/之/此れ)」であることがそれぞれ(　)で括って示されている．ya は並立助詞「や」，de は格助詞「で」を意味する．

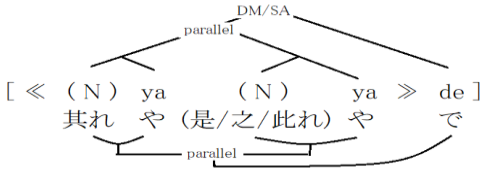

図 2「其れ-や-(是/之/此れ)-や-で」の構造

Figure 2　Syntactic structure denoted by [《(N)ya(N)ya》de] given to「其れ-や-(是/之/此れ)-や-で」

いわゆる CFG における句構造をなしていない表現，すなわち**不完全句**が日本語に見られる特異性の一つであるが，本レキシコンではこの種の表現の構造記述に，欠落した要素の位置を示す空要素記号(null constituent symbol)「$」を用いている．本レキシコンに現れる不完全句には，4.5 の 1, 2 に述べた右開放型と左開放型が有る．例えば，右開放型不完全句「冗談-は-兎_も_角」に与えた構造　[[Nha][D$]]を図 3 に示す．

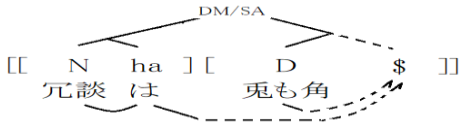

図 3「冗談-は-兎_も_角」の構造

Figure3　Syntactic structure denoted by [[Nha][D$]] given to「冗談-は-兎_も_角」

この表現は，本来，点線で完結する潜在的な句構造をもっており，係り先が空のまま慣用されるに至ったものと考えられる．

いっぽう，左開放型不完全句「と-は-(言/云/謂)う-ものの」には，構造記述[[[[$to]ha]V30]monono]によって図 4 の構造を与える．

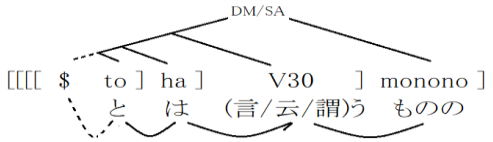

図 4「と-は-(言/云/謂)う-ものの」の構造

Figure 4　Syntactic structure denoted by [[[[$to]ha]V30]monono] given to 「と-は-(言/云/謂)う-ものの」

この場合も点線で示す様な潜在的な句構造の係り元が空のまま先行する文や段落を暗黙裡に$位置に想定した表現と考えられる．

4.7 内部修飾可能性(ギャップ付き構造記述)

F 欄の構造記述内には，例えば図 1 の如く，必要に応じてアステリスク「*」が含まれている．アステリスクは，直後の句の主辞に対する修飾句がこの位置に入り得ることを意味する．従って，図 1 の「斯う−し−た−状況−の−下」の構造記述[[[[[Dsi]ta]*N]no]N]は，図 5 の如く，「斯う−し−た−危険−な−状況−の−下」など，「状況」N の前に種々の連体修飾句が挿入された表現も談話指標として可能であることを意味する．

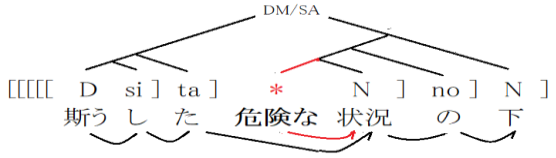

図 5 「*」によって拡張された表現「斯う−し−た−危険−な−状況−の−下」の構造

Figure 5 Syntactic structure of the expression「斯う−し−た−危険−な−状況−の−下」extended by「*」in [[[[Dsi]ta]*N]no]N

JMWEL のこのようなギャップ付き構造記述は定型表現に残っている柔軟性を留保するのに有効である． 修飾句の入り得る場所は係り受けの非交差条件によって決まる．すなわち，P を句とするとき，*P は，P(の主辞)に対する修飾句が，P の直前，および P と主辞を共有し，かつ P を包含する句の直前に存在できることを意味する．従って，P_i (1≦i≦n)が句のとき，右分岐構造の表示 P_1 [P_2 [P_3 [...[P_{n-1} * P_n]...]]]は P_1* [P_2* [P_3 *[...* [P_{n-1} * P_n]...]]]と等価で，P_n (の主辞)に対する修飾句は P_1* [P_2* [P_3 *[...* [P_{n-1} * P_n]...]]]の各アスタリスク位置に存在可能である．

4.8 前方文脈条件(G 欄)

本レキシコンの表現は文頭に置かれて話の流れを制御する場合が多く，この事を G 欄に <top of sentence> と記載している．文頭に来ることが必須の場合は，<top of sentence>* などとアスタリスク「*」を付している．

4.9 後方文脈条件(H 欄)

本レキシコンの表現は，直後に読点を配して使われることが多く，この事を H 欄に<punct. concat.>と記載している．また，例えば，「何故かと言えば…地球が温暖化しているからである」のように，一部の表現は文末側に呼応表現を要求する．これらの呼応表現は， <「からだ」,「からである」,「せいだ」,「せいである」,「ためだ」,「ためである」>*, <「か」,「とは」,「なんて」,「のだが」,「のに」>* , <「そうだ」,「とのことだ」,「らしい」,「という」>などと直接，語句を記載している．

4.10 備考(I 欄)

一部の表現に言い換え情報を与えている．

5. むすび

自然言語処理(NLP)，特に談話処理で重要な点の一つは，文書・談話中にみられる，論述(文，段落)の内容を対象化して論述する，いわば高階論理的な構造を如何に捉えるかである．日本語の場合，そのような構造の手掛かりを一義的に与えてくれるのが文頭の接続表現(前方，後方照応)部分と文末のいわゆる文末表現(前方照応)部分である．ただし，文頭の接続詞，間投詞，副詞，文末の助動詞，終助詞などの単語だけでは現実の文書，談話処理では明らかに不十分であり，複単語表現 MWE を考慮することが不可欠である．本論文では文頭の談話指標・文修飾的 MWE 群について述べた．因みに，文末表現については筆者らの日本語複単語表現レキシコン JMWEL _post-predicative に約 5,000 種が収録・整理されている[6]．これらのレキシコンがこれからの日本語処理，日本語研究進化の一助となれば幸いである．

参考文献

[1] D. Biber, S. Johansson, G. Leech, S, Conrad, E. Finegan (eds.). Longman Grammar of Spoken and written English. Harlow: Pearson Education Limited, 1999.
[2] R. Corrigan, E. A. Moravcsik, H. Ouali, K. Wheatley (eds.). Formulaic Language Vol.1, Distribution and historical change. John Benjamins Publishing Company, 2009.
[3] 藤井聖子. “条件構文の談話指標化の諸相”. 第 4 回コーパス日本語学ワークショップ予稿集，国立国語研究所， 2013.
[4] Joshi, A. Multiword Expressions as Discourse Relation Markers (DRMS). invited talk, Workshop on Multiword Expressions at COLING2010, 2010.
[5] I.A.Sag, T. Baldwin, F. Bond, A. Copestake, D. Flickinger. Multiword Expressions-A Pain in the Neck for NLP. Proc. of the 3rd CICLING, 2002.
[6] 首藤公昭. “日本語処理研究工房 ことばの森”. http://jefi.info, 2011 開設.
[7] T. Tanabe, M. Takahashi, K. Shudo. A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing. Computer Speech and Language, 28-6, Elsevier, 2014.
[8] "Joint Workshop on Multiword Expressions and Wordnet (MWE-WN-2019)". http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_03_MWE-WN-2019___lb__ACL__rb__ , 2019.