正規表現による文型検索ツールの試作――IPADic と UniDic の利用をめぐって

蔡佩青（淡江大学外国語学部日本語文学科）
魏世杰（淡江大学商学学部情報経営学科）

A Sentence Pattern Filter Based on Regular Expressions: Using IPADic and UniDic

Tsai, Pei-Ching (Department of Japanese, Tamkang University)
Wei, Shih-Chieh (Department of Information Management, Tamkang University)

要旨

発表者は,日本語学習者が文章を作成する際の文型応用力を高めるための,文型検索ツールの開発を提案したことがある[1]。すなわち,入力した文章には指定の文型に当てはまる文があれば,自動的にリストアップされ文型部分が赤字で表示されるような,文型検索のユーザインタフェースを構築するのである。検索ツールは正規表現（Regular Expression）をもってプログラミングするが,形態素解析は MeCab を,辞書は IPADic を用いた。ところが,試作版のテスティングでは,MeCab の誤解析による文と文型とのミスマッチングが起こった。その一部は IPADic の形態素に付与する品詞情報に起因すると考えられる。そこで,IPADic を UniDic に替えることで殆どのミスマッチングが解消できた。本稿は,二回にわたって行ったテストと成果のまとめ報告である。

１. はじめに

言語習得の四技能とされる「聞く・話す・読む・書く」の中で,上達が最も遅れるのは書くことだと言われている。サンプル数は多くないが,発表者は勤務校で日本語学科二年次の作文授業を担当した時, 800 字程度の学生の自己紹介文にどのような文型が取り上げられたかを集計したことがある。結果,ほぼ一年次前期で学習した文型しか使われなかったのだ。テーマの設定もアンケート結果に影響を与えているだろうが,学習済みの文型を上手に作文に活かせるためには,文章における文型や語法の使い方,つまり文脈に沿った文型の選び方が重要だと考える。そこで,正規表現（Regular Expression）[2]を用いて,学習者の文型応用力を高めるための,文型検索ツールの開発を提案した。本稿では, 蔡・魏(2020)で報告した,文型検索ツール試作版の開発過程をまとめしつつ,試作版で残した課題の解決方法を探る。

２. パターンマッチングテスト

検索ツールの作成は文型の選定,テスト用例文の作成,そしてプログラムの設計・構築と

---

[1] 蔡佩青・魏世杰(2020)「正規表現による文型検索ツールの提案と試作」。
[2] メタ文字（Meta Character）と呼ばれる特殊文字を用いて,文字列を一つの形式で表現する方法である。詳しくは https://docs.python.org/3/library/re.html を参照されたい。

いったステップをふんで進めていくが,それに先立って正規表現による文型のパターンマッチング（Pattern matching）が正確にできるか否かのテストをまず行った。テストの流れは次の通りである。

①文型を 20 選出する。

②各文型に対して例文を 10 作成し,計 200 文をパターンマッチングテスト用のデータにする。

③文型の正規表現を作成して, ②で作成した 200 の例文をもってパターンマッチングテストを行う。

２．１文型の選定

文型検索ツールは初級・中級日本語学習者の利用を想定して開発するものであるが,「A は N だ」「A は V する」のような,ほぼすべての文に当てはまってしまう基本文型を除いて,なるべく用言の活用を含む文型を選り出した。テストに使用した文型の基本型は表 1 の通りである。また,すべての文型における肯否,テンスや助動詞の活用形が含まれている。

表 1 パターンマッチングテスト用文型

1 動詞未然形＋ない＋て＋は＋いけない	11 形容詞連用形（く）＋動詞
2 動詞連用形＋て＋いる	12 形容詞連用形＋たり
3 動詞連用形＋た＋こと＋が＋ある	13 形容詞連体形＋名詞
4 動詞連用形＋た＋ところ	14 形容詞仮定形＋ば
5 動詞連用形＋たい／たがる	15 形容詞語幹＋そうだ
6 動詞終止形＋と	16 形容動詞連用形（に）＋動詞
7 動詞連体形＋こと＋が＋できる	17 形容動詞連用形＋たり
8 動詞連体形＋ようだ	18 形容動詞連体形＋名詞
9 動詞仮定形＋ば＋動詞連体形＋ほど	19 形容動詞仮定形（＋ば）
10 動詞意志形＋思う	20 形容動詞語幹＋そうだ

上記の文型には,学問分野によっては文法項目,もしくは定型表現といった部類に入るものも含まれているかもしれない。しかし,外国人日本語学習者がよく使う市販の日本語教材において,その殆どは単語→文型→例文のように,学習目標となる文法や表現を「文型」として例文とともに挙げている。また,日本語教師も多く参考にしている『教師と学習者のための日本語文型辞典』もかなり広義的に「文や節の意味,機能,用法にかかわる形式」を文型としている[3]。そこで,学習者に混乱を起こさせないために,本研究においても,類型として学習者が習得すべきと考えられる作文上の文法や語法を含めて文型とする。

２．２例文の作成

表 1 の文型をもって,それぞれ例文を 10 文作成した[4]。実際の文章における応用を考慮し,

---

[3] グループ・ジャマシイ編(1998)『教師と学習者のための日本語文型辞典』くろしお出版。

[4] 例文の一部は, 蔡佩青(2019)『修訂新版日本語句型知惠袋』を参考にして作成した。

構文が少々複雑で長い文を多く作るようにしている。そのため,一つの例文に複数の文型が含まれることもある。例えば,「12 形容詞連用形＋たり」の例文として,「年末になると,夫は仕事が忙しかったり,クリスマスや忘年会などイベントが続いたりして,いつも帰りが遅くなる」を採り上げたが,「6 動詞終止形＋と」と「11 形容詞連用形（く）＋動詞」の文型も使われている。

２．３正規表現の作成

文型は様々な単語を入れ替えたりして文を作るためのもので,文型によっては単語に活用形や音便などの制限がある。そのため,単語に品詞・品詞細分類・活用型・活用形などの情報（素性）が付されている日本語形態素解析システム MeCab を活用して,文型の正規表現を作成していく。形態素解析後の出力フォーマットと正規表現の記述とは,表 2 のように対応している。

表 2　MeCab の品詞付けと正規表現の記述との対応

MeCab	表層形	発音	原形	品詞（細分類 1・2・3 を含む）	活用型	活用形
正規表現	word	kana	lemma	pos	itype	iform

作成した文型の正規表現は次のようである。例として,動詞・形容詞・形容動詞の活用をそれぞれ含むものを一つずつ掲げる。

4 動詞連用形＋た＋ところ
<word:た,kana:タ,lemma:た,pos:助動詞,itype:特殊・タ,iform:基本形>
<word:ところ,kana:トコロ,lemma:ところ,pos:名詞-非自立-副詞可能,itype:,iform:>

12 形容詞連用形＋たり
<word:.*,kana:.*,lemma:.*,pos:形容詞-自立,itype:形容詞.*,iform:連用タ接続>
<word:たり,kana:タリ,lemma:たり,pos:助詞-並立助詞,itype:,iform:>

19 形容動詞仮定形（＋ば）
<word:.*,kana:.*,lemma:.*,pos:名詞-形容動詞語幹,itype:,iform:>
?(<word:なら,kana:ナラ,lemma:だ,pos:助動詞,itype:特殊・ダ,iform:仮定形>|
?<word:で,kana:デ,lemma:だ,pos:助動詞,itype:特殊・ダ,iform:連用形>
<word:なけれ,kana:ナケレ,lemma:ない,pos:助動詞,itype:特殊・ナイ,iform:仮定形>
<word:ば,kana:バ,lemma:ば,pos:助詞-接続助詞,itype:,iform:>)

３．パターンマッチングテストの結果

テストでは 200 の例文をもって 10 の文型とパターンマッチングを行った。理論上,文型に従って例文を作成したので,例文はそれぞれの文型にマッチするはずである。ところが,テストの結果,例文に含まれている文型にマッチしなかった文（F 文とする）と,マッチするはずのない文型にマッチしてしまった文（U 文とする）が存在していることが分かった。

３.１MeCab 誤解析の問題

次に掲げるのは文型 5 と文型 6 の例文（下線は筆者,以下同）だが,いずれもパターンマッチングテストで F と判定されている。

5 動詞連用形＋たい／たがる
F:あのスーパーは会員向けの無料配送サービスを開始したので,利用してみたいです。
6 動詞終止形＋と
F:外国の人にも読んでもらいたかったら,英語で書くといいです。

そこで,F とされる理由を探るべく,形態素解析の詳細を確認してみたところ,MeCab の誤解析と辞書の品詞設定が原因であることが判明した。

文型 5 は動詞の連用形に希望を表す助動詞「たい」を接続するものであるが,F とれる例文の下線部「みたい」は,「見る」と「たい」との組み合わせではなく,一つの単語,すなわち比況表現の「みたい」と見なされたため,マッチすることはできなかった。また,文型 6 について,「書くといい」の「と」は,形態素解析では「助詞-格助詞-引用」の「と」として解析され,文型に取りあげた接続助詞の「と」と異なるものとなっている。そのためか,後続語の「いい」は形容詞ではなく動詞の「言う」の連用形と解析されてしまう。

MeCab の品詞付けの誤りについて,未知語（辞書にない形態素）の存在が大きな要因としてしばしば言及されている[5]。ことに新語や流行語が多く含まれている SNS 情報を解析する際の誤解析がよく発生するという。また,表記の揺れ・オノマトペ・連濁・方言・長音化なども誤りが起こりやすい原因とされている[6]。しかし,上記の問題はそれらが原因ではないようだ。

３.２IPADic の品詞付けの問題

解析の精度を上げるために, MeCab 用の修正プログラムや拡張辞書など多く存在している[7]が,今回のテストでは IPADic[8]を使用した。IPADic は MeCab の初期設定で推奨される辞書であって,日本語文法に似た品詞体系を持っているため,文型検索ツールの構築に適していると考えた。ところが,電子計算機の情報処理のメカニズムに合わせるためか,IPADic には国文法や日本語文法で定められている品詞分解の法則とは異なるものがある。今回のテストで気づかされたのは形容動詞の問題である。

20 形容動詞語幹＋そうだ
F:言い負かされた男は不満そうに速足で店を出た。

上に掲げた文型 20 の F 文について,「不満」という語は一般的に形容動詞として認められているにも関わらず[9],「形容動詞語幹＋そうだ」の文型に当てはまらず,パターンマッチ

---

[5] 中村純平・伝康晴(2008)「形態素解析誤りの多い助詞・助動詞の再解析」,小山照夫・竹内孔一(2015)「形態素解析の系統的誤りと用語抽出」他。
[6] 鍜治伸裕他(2015)「形態素のエラー分析」。
[7] Web 上の新語・流行語が追加・更新できる MeCab 対応の拡張辞書「NEologd」が有名である。
[8] IPA コーパスに基づき CRF でパラメータ推定した辞書である。https://taku910.github.io/mecab/
[9] 北原保雄編(2010)『明鏡国語辞典』第二版。

ングテストに通らなかった。下線部の「不満そうに」を MeCab で形態素解析すると,その品詞情報は次の通りになる（「-」は品詞の細分類を表す。以下同)。「不満」は「名詞」として扱われているので F に判定されたことが判明した。

不満：名詞-一般
そう：名詞-接尾-助動詞語幹
に　：助詞-副詞化

そして,同じ辞書設定の問題で,今度は U となった文がある。

16 形容動詞連用形（に）＋動詞
U:疑問に感じたことがあります。

上記の文は,経験を表す文型「3 動詞連用形＋た＋こと＋が＋ある」の例文として採り上げたが,文型 16 にもマッチしている。しかし,この文は 2.2 節で説明したような,複数の文型に当てはまることを想定して作成した例文ではない。筆者は「疑問」という語を名詞として考えて[10]例文を作ったため,形容動詞の文型にマッチしたことに少々困惑する。MeCab の形態素解析を確認すると,「疑問」を単なる名詞ではなく,品詞再分類で形容動詞語幹ともタグ付けされている。確かに「疑問に思う」や「疑問に感じる」などのように,知覚動詞とともに表現すると,形容動詞に似たニュアンスが表れる。そもそも形容動詞の品詞問題について,文法学説においてもよく議論され,研究者によって主張が異なっている。しかし,文型の正規表現を作成するにあたって,こういった特殊な例を一つひとつ取り上げて処理するのには多くの時間を要することになろう。

４. ユーザインタフェースの構築

3 節で述べた形態素解析上の問題を残しつつ,文型検索のユーザインタフェースを構築してみた。インタフェースに任意の文章を入力すると,表 1 に挙げた 20 の文型に当てはまる文があれば,自動的にリストアップされると同時に,指定の文型が使用されている箇所が赤字で表示されるようになる。図 1 はその試作版のテスト画面である。なお,図 1 に使用した文章は本稿 2.1 節の一部である。

[10] 注 9 の辞書によって確認している。

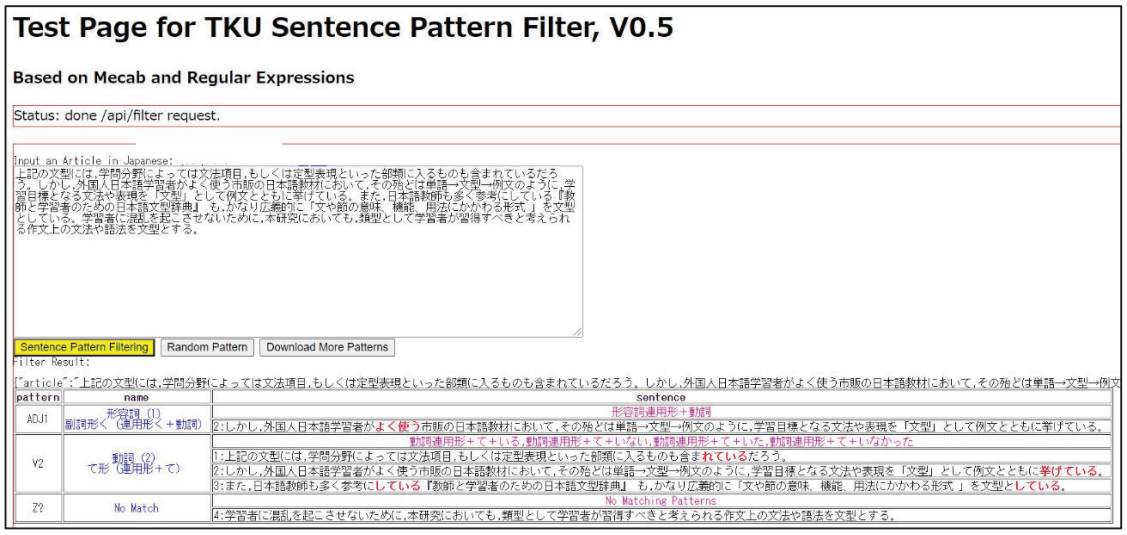

図 1　文献検索ルーツ試作版 20200621V0.5

５.おわりに代えて―UniDic による形態素の品詞付け

以上は蔡・魏(2020)で報告した研究成果である。その後,MeCab の誤解析と辞書の品詞付けに関する問題を解決すべく,まず拡張辞書を UniDic に換えて検討してみた。使用したのは現代書き言葉 UniDic[11]である。

3 節で挙げた F 文と U 文を,UniDic を実装した MeCab で解析してみると,品詞付けはすべて例文の設定通りになった。つまり,「利用してみたい」の「たい」は助動詞,「書くといい」の「いい」は形容詞,「疑問に感じた」の「疑問」は名詞とされているのである。そして,IPADic では,名詞とされているためにパターンマッチングテストに通らなかった「不満」については, UniDic では「名詞-普通名詞-形状詞可能」というように品詞分類されている。ここでいう形状詞は形容動詞のことを指している[12]。日本語学習教材に見られる形容動詞やナ形容詞という名称は使われていないが,正規式を作成する際に形容動詞と形状詞を入れ替えれば問題なかろう。

上述したように,IPADic を UniDic に換えたことで,蔡・魏(2020)における課題はほぼ解決できたように思う。今後, UniDic の品詞体系に合わせた正規表現を新たに作成していき, IPADic と UniDic による文型のパターンマッチングテストの結果を比較するとともに,学習者が実際に文型検索インターフェースを使用して学習する状況についてのアンケート調査も行う予定である。

参考文献

鍛治伸裕・森信介・高橋文彦・笹田鉄朗・斉藤いつみ・服部圭悟・村脇有吾・内海慶(2015).「形態素のエラー分析」言語処理学会第 21 回年次大会ワークショップ.(https://www.anlp.jp/proceedings/annual_meeting/2015/html/paper/WS_PNN02_morphological-analysis.pdf より

[11] 拡張辞書は https://unidic.ninjal.ac.jp/download#unidic_bc,パッケージのソースコードは https://github.com/polm/fugashi より。

[12] UniDic の品詞体系によると,「『静か』『健やか』など、いわゆる形容動詞の語幹部分」を「形状詞」とし,「名詞としての用法があるものは、『名詞-普通名詞-形状詞可能』」に分類しているという（伝康晴・山田篤・小椋秀樹ほか(2008)『UniDic version 1.3.9 ユーザーズマニュアル』)。

ダウンロード可能)
北原保雄編(2010).『明鏡国語辞典』第二版,大修館書店.
グループ・ジャマシイ編(1998).『教師と学習者のための日本語文型辞典』くろしお出版.
小山照夫・竹内孔一(2015).「形態素解析の系統的誤りと用語抽出」情報処理学会研究報告,pp.1-4.(http://research.nii.ac.jp/~koyama/official/tmdb/pdf/correct.pdf よりダウンロード可能)
蔡佩青(2019).『修訂新版日本語句型知惠袋』眾文圖書股份有限公司.
蔡佩青・魏世杰(2020).「正規表現による文型検索ツールの提案と試作」『AI と日本語教育との協働』国際シンポジウム会議予稿集,pp-92-99.
中村純平・伝康晴(2008).「形態素解析誤りの多い助詞・助動詞の再解析」言語処理学会第14回年次大会発表論文集,pp.73-76.
伝康晴・山田篤・小椋秀樹・小磯花絵・小木曽智信(2008).『UniDic version 1.3.9 ユーザーズマニュアル』p.16.(https://unidic.ninjal.ac.jp/UNIDIC_manual.pdf よりダウンロード可能)

関連 URL

Python regular expression documentation: https://docs.python.org/3/library/re.html

MeCab with IPADic: https://pypi.org/project/mecab-python3/0.996.5/

MeCab with UniDic: https://github.com/polm/fugashi