

国立国語研究所学術情報リポジトリ

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

メタデータ	言語: Japanese 出版者: 公開日: 2021-03-05 キーワード (Ja): キーワード (En): IPADic, UniDic 作成者: 蔡, 佩青, 魏, 世杰, Tsai, Pei-Ching, Wei, Shih-Chieh メールアドレス: 所属:
URL	https://doi.org/10.15084/00003153

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（淡江大学商管学部情報経営学科）

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い。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。蔡・魏(2020)¹ではその試作版の開発について報告した。検索ツールは正規表現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、テストでは、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 に替えることで多く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文型パターンマッチングテストの結果と検討事項をまとめた報告である。

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 2 年次の作文授業を担当した時、800 字程度の自己紹介文で学生がどのような文型が取り上げているかを集計したことがある。結果、ほぼ 1 年次前期で学習した文型しか使わなかったことが分かった。テーマの設定も集計の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、文章読解に必要な文型判別力及び文章作成上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い。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つつ、試作版で残した課題の解決方法を探る。

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。

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,正規表現の作成,そしてプログラムの設計・構築と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)が正確に実行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 文作成し,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,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」³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き,なるべく用言の活用を含む文型を選び出した。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定,否定,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(かつ)+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(だっ)+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は,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行っている『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⁴。そこで,学習者に混乱を起こさないために,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。

³ 「A は N だ」は名詞文を,「A は V する」は動詞文を指している。

⁴ 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。

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⁵。実際の文章における応用を考慮し、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・活用型・活用形などの情報が付され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のように対応させる。

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なものである。例として、動詞・形容詞・形容動詞の活用がそれぞれ含まれる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

正規表現の作成が完成したのち、前述した200の例文をもって10の文型とパターンマッチングを行う。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

⁵ 例文の一部は、蔡佩青(2019)『修訂新版日本語句型知恵袋』を参考にして作成した。

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。

3.1 MeCab 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 文と判定されている。

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 文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 と判定された例文の下線部「みたい」は,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,形態素解析では引用の格助詞として解析され,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」の連用形である「言い」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⁶。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⁷。しかし,上記の問題はそれらが原因ではないようだ。

3.2 IPADic の品詞付けの問題

解析の精度が上がるための, 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⁸が,今回のテストでは IPADic⁹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。

⁶ 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。

⁷ 鍛冶伸裕他(2015)「形態素のエラー分析」。

⁸ 例えば, 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。

⁹ 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/

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¹⁰、「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ために F 文に判定されたようだ。

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 文となった例文がある。

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¹¹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、品詞再分類では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は多くの時間を要することになるので、あまり現実的ではない。

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 北原保雄編(2010)『明鏡国語辞典』第二版。

¹¹ 注 10 の辞書によって確認している。

Test Page for TKU Sentence Pattern Filter, V0.5

Based on Mecab and Regular Expressions

Status: done /api/filter request.

Input an Article in Japanese:

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だろう。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、その殆どは単語一文型一例文のように、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、学習者に法則を教へさせるために、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文型とする。

Sentence Pattern Filtering Random Pattern Download More Patterns

Filter Result:

["article"] 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だろう。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、その殆どは単語一文型一例文のように、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。

pattern	name	sentence
ADJ1	形容詞 (1) 副詞形+く(連用形+て+動詞)	2: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、その殆どは単語一文型一例文のように、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。
V2	動詞 (2) て形(連用形+て)	1: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だろう。 2: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、その殆どは単語一文型一例文のように、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。 3:また、日本語教師も多く参考している『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。
Z?	No Match	No Matching Patterns 4:学習者に混乱を起こさないために、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文型とする。

図1 文献検索ルーツ試作版 V0.5

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の概要である。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 UniDic¹²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」は名詞という品詞情報が付されているのである。そして、IPADic では、名詞とし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¹³。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、正規表現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)でぶつかった課題はほぼ解決できたように思う。ただ、一部の形態素誤解析の問題は依然として残っている。例えば、文型 5 の例文として挙げた「彼女も今回の交換留学に参加したがっています」について、「参加したがっています」の部分は、IPADic を使用しても UniDic を使用しても「参加／したがう／て／いる」のように品詞分解をされてしまう。今後、残された課題の解決策を探りつつ文型を増やして、異なる辞書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。

最後に、蔡・魏(2020)以来さらにバージョンアップをした文型検索ユーザインタフェース試作版の画面を掲げておく (図 2)。辞書の IPADic と UniDic が選択できるようにすると

¹² 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。

¹³ 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』)。

もに、各文型の選択ボタンも増設した。画面中の文章は図 1 と同様なものであるが、辞書は UniDic を選択している。

Test Page for TKU Sentence Pattern Filter, 日文句型分類儀, V0.6

Based on Mecab and Regular Expressions

Status: done /api/filter request.

上記の文型には、学問分野によっては文法項目、もしくは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は、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」している『教師と学習者のための日本語文法辞典』もかなり体系的に「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。そこで、学習者に混乱を起させないために、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。

Dict_Type 断詞字典: ☐ IPADic ☒ UniDic
 Sentence Pattern 分類句型:
☒ V1 ☒ V2 ☒ V3 ☒ V4 ☒ V5 ☒ V6 ☒ V7 ☒ V8 ☒ V9 ☒ V10
☒ ADJ1 ☒ ADJ2 ☒ ADJ3 ☒ ADJ4 ☒ ADJ5 ☒ ADJV1 ☒ ADJV2 ☒ ADJV3 ☒ ADJV4 ☒ ADJV5

Filter Result 分類結果:

"article"="上記の文型には、学問分野によっては文法項目、もしくは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は、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ます。また、日本語教師も多く「参考」している『教師と学習者のための日本語文法辞典』もかなり体系的に「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。そこで、学習者に混乱を起させないために、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。	pattern	name	sentence (dict_type=unidic)
ADJ1	副詞(短く)連用形(＋動詞)	形容詞(1) 形動(短く)連用形(＋動詞)	1: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は、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ます。 動詞連用形＋ている, 動詞連用形＋ていない, 動詞連用形＋ていた, 動詞連用形＋ていなかった
V2	動詞(2) て形(連用形＋て)	上記の文型には、学問分野によっては文法項目、もしくは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。2: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は、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ます。3:また、日本語教師も多く「参考」している『教師と学習者のための日本語文法辞典』もかなり体系的に「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。	
Z?	No Match	No Matching Patterns	3:そこで、学習者に混乱を起させないために、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。

図2 文献検索ルーツ試作版 V0.6

参考文献

- 鍛冶伸裕・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恵袋』 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・小木曾智信(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>