データの概要とタグの検索

統語・意味解析コーパス (NPCMJ) チュートリアル

吉本啓・長崎郁

2019.5.11

はじめに

- NPCMJ ウェブサイト（`http://npcmj.ninjal.ac.jp/`）に従って，以下について紹介する。
  - 概要とコンテキスト表示，分析表示，タグ（タグ・ブラウザー）

概要とコンテキスト表示

- トップページの NPCMJ ツール - NPCMJ Search を開く をクリック
- 検索インターフェース一覧へ
- 検索インターフェース一覧の最上段左の画像から「概要・コンテクスト表示」へ

概要とコンテキスト表示（2019.3.8 現在）

Source	ツリー数	語数	
aozora	4646	101537	青空文庫
bible	1664	30657	聖書
book	552	12515	様々な書籍の原文の一部
dict	3419	33651	辞書の例文
diet	1698	37349	国会議事録
fiction	923	12051	小説の抜粋
law	337	7793	法律文
misc	2085	23872	その他
news	4666	84927	新聞記事
nonfiction	223	4454	手紙など
ted	1453	22030	TED 日本語
textbook	6048	64038	文法書の例文
wikipedia	2746	70445	ウィキペディア記事
Total	30460	505319	

「概要」ページ

- データ（小説（青空文庫）など）を拡充し，プロジェクト期間内に 6 万文の規模となることを目指している。
- Download all bracketed trees から括弧付きツリー形式のすべてのデータ（kana version と romanised version）をダウンロード可能
- 収録テキスト一覧
  - 各テキストのへのアクセス
- テキスト番号をクリックし，「コンテキスト表示」へ

コンテキスト表示

- テキストのメタデータ
- テキスト全文
- ローマ字表記（右上の ABC）と漢字仮名交じり（伊呂波）の切り替え
- 表示されたテキストの「括弧付きツリー」データのダウンロード
  - 括弧付きツリーをダウンロード をクリック）
- 文番号をクリックし，「分析表示」へ

分析表示

- 統語・意味解析の結果をツリーの形で表示
- ツリーの移動
  - 前 後 で前後の文へ
- 表示ツリーの追加
  - ツリー左上の矢印（⇑）・左下の矢印（⇓）をクリック
- 表示ツリーの削除
  - ツリー右上の矢印（⇓）・右下の矢印（⇑）をクリック
- 表示された文のデータのダウンロード
  - SVG（表示されているツリーの画像），括弧付きツリー，XML
  - 括弧付きツリーと XML は注釈の情報量が違う。

括弧付きツリー形式（Penn Treebank と同様の手法を用いたもの）

```
( (IP-MAT (PP (NP (N (WORD 初め)))
                  (P-ROLE (WORD に)))
            (PP-SBJ (NP (N (WORD 言)))
                    (P-ROLE (WORD が)))
            (VB (WORD あっ))
            (AXD (WORD た))
            (PU (WORD 。)))
  (ID 1_bible_new;b.JOH.1.1;JP))
```

- 括弧付きツリー形式では，各要素に対して，WORD ノードを介して品詞ラベル（N，P，VB，AXD）が，さらに品詞ラベルを土台に句・節ラベル（NP，PP，IP）が付与されている。
- ただし，WORD ノードは例文のツリー表示には反映されない。また，本ウェブサイトの検索インターフェース（TGrep-lite）で検索する際にも無視される。
- 品詞ラベルも句・節ラベルも，拡張ラベル（ハイフンの後）によって下位類や機能の情報を与えられる。

XML 形式

```
<alpino_ds id="1_bible_new;b.JOH.1.1;JP" version="1.3">
  <node cat="ip-mat" id="1" begin="0" end="7">
    <node cat="pp" id="2" begin="0" end="2">
      <node cat="np" id="3" begin="0" end="1">
        <node cat="n" lemma="初め" aform="ハジメ" pos="名詞-普通名詞-副詞可能" pron="ハジメ" id="4" begin="0" end="1">
          <node pt="word" word="初め" id="5" begin="0" end="1">
          </node>
        </node>
      </node>
      <node cat="p-role" lemma="に" aform="ニ" pos="助詞-格助詞" pron="ニ" id="6" begin="1" end="2">
        <node pt="word" word="に" id="7" begin="1" end="2">
        </node>
      </node>
    </node>
    <node cat="pp-sbj" id="8" begin="2" end="4">
      <node cat="np" id="9" begin="2" end="3">
        <node cat="n" lemma="言" aform="ゲン" pos="名詞-普通名詞-一般" pron="ゲン" id="10" begin="2" end="3">
          <node pt="word" word="言" id="11" begin="2" end="3">
          </node>
        </node>
      </node>
      <node cat="p-role" lemma="が" aform="ガ" pos="助詞-格助詞" pron="ガ" id="12" begin="3" end="4">
        <node pt="word" word="が" id="13" begin="3" end="4">
        </node>
      </node>
    </node>
    <node cat="vb" lemma="有る" aform="アル" pos="動詞-非自立可能" pron="アッ" stype="五段-ラ行" vform="連用形-促音便" id="14" begin="
      <node pt="word" word="あっ" id="15" begin="4" end="5">
      </node>
    </node>
    <node cat="axd" lemma="た" aform="タ" pos="助動詞" pron="タ" stype="助動詞-タ" vform="終止形-一般" id="16" begin="5" end="6">
      <node pt="word" word="た" id="17" begin="5" end="6">
      </node>
    </node>
    <node cat="pu" lemma="。" pos="補助記号-句点" id="18" begin="6" end="7">
      <node pt="word" word="。" id="19" begin="6" end="7">
      </node>
    </node>
  </node>
  <sentence>初め に 言 が あっ た 。</sentence>
</alpino_ds>
```


- XML 形式では，lemma（見出し語），pos（UNIDIC による品詞ラベル），begin と end（ノードの線形順）などの情報が付与されている。

データ形式について

- 括弧付きツリー形式
  - Penn Treebank のアノテーション方法を日本語に改変・適用したもの
  - 全データをこの形式でダウンロードできる（→「概要」ページ）
  - 本ウェブサイトの検索インターフェース（TGrep-lite）で検索が可能
  - Tregex（`https://nlp.stanford.edu/software/tregex.shtml`）などを利用したローカル環境での検索が可能
- XML 形式
  - Alpino XML（`http://www.let.rug.nl/vannoord/trees/`）による
  - 本ウェブサイトの検索インターフェース（XPath）で検索が可能

分析表示（続）

- 表示モードの切り替え
  - default（デフォルト表示）　→ 最初に表示されるもの
  - indexed（インデックス表示）
  - dependency（依存関係表示）
  - formula（計算式表示）

分析表示（続）

- default（デフォルト表示）

  品詞ラベル横の短剣符にカーソルを合わせると，lemma 情報と pos 情報（いずれも UNIDIC に従ったもの）を確認することができる。

分析表示（続）

- indexed（インデックス表示）

- ツリーに項構造情報と照応関係（big PRO を含む）を示すインデックス情報が表示される。
  - アノテーションでは明示されない文法的依存関係が明示されるようになる。
  - (IP-EMB　悶々とした) の (ZERO　*PRO*) に注目
  - コントロール環境（当該の節の主語が上位の節の項から継承される環境）や ATB を定義しているため，ゼロ代名詞が自動的に付与される。
- ノード上の赤い番号 → 述語に対する項や修飾要素となるノードに付与
- 述語の下の茶色い番号 → 述語のとる項や修飾要素を示す
- グレーの点線 → 項／修飾と述語とを結ぶ

分析表示（続）

- dependency（依存関係表示）

- ターゲット（灰色）と句（オレンジ）の関係を表示
- ターゲットごとに，他の句との関係が示される。
- ターゲットと句の関係は，句の枠の先頭に示される（ARG0、ARG1、REL、等）。

分析表示（続）

- formula（計算式表示）
  - 述語論理に基づいた式が意味表現として示される（意味表現はツリーにおける文法関係から生成）
  - 画面上部の TPCP をクリックすると，定理証明器（Sutcliffe 2009）にかけることのできる式を生成
  - 文法関係から意味表現に至る変換の過程を見ることもできる

    [0] — 出発点となるアノテーションを表示
    [1] — normalize したツリーを表示
    [2] — 上記を Prolog に変換したものを表示
    [3] — 上記を SCT 表現に変換したものを表示
    [4] — 最初の意味計算の結果を表示
    [5] — 再位置決め編集後の結果を表示
    [Formula] ボタンをクリックすると、計算式表示に戻る

タグ（タグ・ブラウザー）

- 「タグ」ページへの入口
  - 検索インターフェース一覧から
    （一覧の上から二番目，左の画像から「タグ」へ）
  - インターフェースの各ページから
    （画面上部メニューの タグ をクリック）

「タグ」ページ

- NPCMJ で使われているタグ（ラベル）と，その具体例をこのページから見ることができる → 「タグ（ラベル）付けの一般規則」へ
- 画面右側の Part-of-speech（品詞），Phrase（句）タグ」をクリック → それぞれのリストが開く（もう一度クリックすると閉じる）。
- タグのボタンをクリックすると検索結果が表示される（検索したタグに該当するハイライトされる）。

「タグ」ページ

- 検索結果
  - 画面上部：
    “TGrep-lite 検索結果”（TGrep-lite という検索言語を使ってタグを検索している）
    その下のボックスは，当該のタグを TGrep-lite 式に表現したもの
  - 画面下部：
    ヒット数，次の検索結果へ（一画面の表示件数は 25 件）
    検索結果は CSV，括弧付きツリー，XML ツリーの形式でダウンロードが可能
  - 各文のデータ番号をクリックすると，ツリーが表示される。

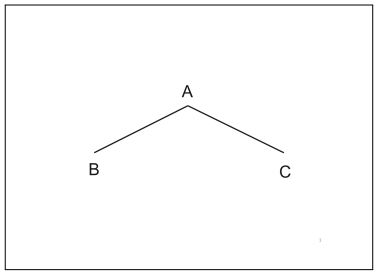
タグ（ラベル）付けの一般規則

カッコ表示と木表示

句構造規則の適用の結果生じる派生の過程（「解析結果」とも呼ぶ）を図示したもの

A ⇒ B C

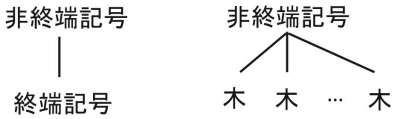
(A
B
C)



どちらでも情報は全く同じ
カッコ表示の行換えは無意味

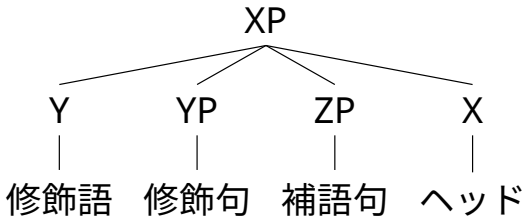
表示の規則

- 木 → (非終端記号　終端記号)
- 木 → (非終端記号　木
  木
  …
  木)



X バー理論

- 抽象的な文法スキーマ：
  - 文法規則を作る規則
  - X は XP を投射（project）する



動詞やイ形容詞の活用形の扱い

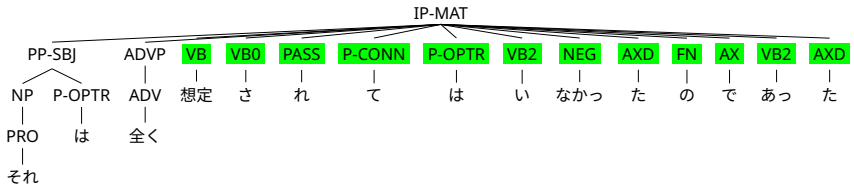
動詞，イ形容詞，助動詞などの活用語に関しては，概ね学校文法的な扱いを採用している

- 動詞（VB）
  - 下一段活用：「食べる」→ 食べる｜食べろ｜食べよ｜食べ｜食べれ
  - 上一段活用：「起きる」→ 起きる｜起きろ｜起きよ｜起き｜起きれ
  - 五段活用：「走る」→ 走る｜走れ｜走ら｜走り｜走ろ
  - カ変活用：「来る」→ 来る｜来い｜来れ｜来
  - サ変活用：「する」→ する｜しろ｜せよ｜すれ｜し

述語の拡張

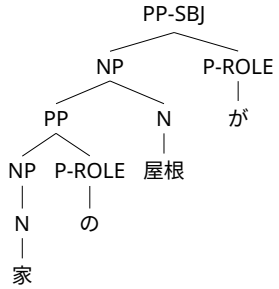
- 拡張された述語の個々の要素は，同一の節（IP）の元にフラットに並べられる

「それは全く [想定されてはいなかったのであった]」



格助詞を伴う名詞句

- 格助詞が主要部（ヘッド）となり，名詞句を補部（complement）として取って，助詞句（PP) を構成



様々なタグ（1）

- タグのタイプには次のようなものがある
  - 「語レベルのタグ」と「句レベルのタグ」
  - 「基盤となるタグ」と「拡張タグ」

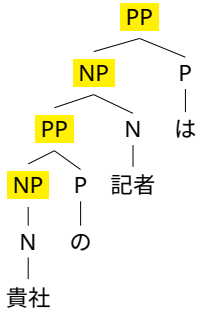
- 語レベルのタグ：
  - 語に（WORD ラベルを介して）与えられる

    N（Noun，名詞），P（Particle，助詞），VB（Verb，動詞），ADJI（I-adjective，イ形容詞），ADV（Adverb，副詞），...

様々なタグ（2）

- 句レベルのタグ：
  - （多くの場合）語レベルのカテゴリーが投射するカテゴリーに対して与えられる

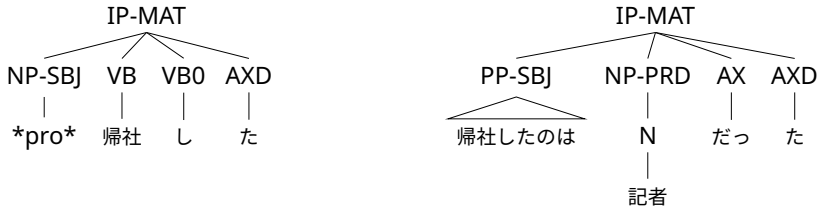
    NP（Noun Phrase，名詞句），PP（Particle Phrase，助詞句），ADVP（Adverbial Phrase，副詞句），IP（Inflectional Phrase，節）...



- NP → N
  (NP (N 貴社))
- PP → NP P
  (PP (NP (N 貴社))
  (P の))
- NP → PP N
  (NP (PP (NP (N 貴社))
  (P の))
  (N 記者))

様々なタグ（3）

- IP（節）を投射するのは述語。
  述語は，VB（動詞），ADJI（イ形容詞），ADJN（ナ形容詞），NP（名詞句），ADVP（副詞句）等の語カテゴリー・句カテゴリーを核として構成される。



pro（3 人称のゼロ代名詞）は，文中に表現されていない主語や目的語を示すために用いられる。

様々なタグ（4）

- 基盤タグ：
  - 語やそれの投射するカテゴリーに対して与えられるタグ
- 拡張タグ：
  - 語または句の下位類や統語的な機能を示すタグ
  - 基盤タグの後にハイフンを付けて付加される

  IP-MAT（Matrix，主節），PP-SBJ（Subject，主語である助詞句），P-ROLE（格助詞），P-OPTR（とりたて助詞），NP-PRD（Predicate，名詞述語）

タグ一覧：品詞タグ I

QUOT　引用符 (quote)
-LRB-　左括弧 (left bracket)
-RRB-　右括弧 (right bracket)
PU　句読点 (punctuation)
ADJI　イ形容詞 (イ-adjective)
ADJI-MD　モーダルなイ形容詞 (modal イ-adjective)
ADJN　ナ形容詞 (ナ-adjective)
ADJN-MD　モーダルなナ形容詞 (modal ナ-adjective)
ADV　副詞 (adverb)
AX　助動詞 (auxiliary verb (including copula))
AXD　テンス標識 (助動詞の一部) (auxiliary verb, past tense), 過去テンス
CL　助数詞 (classifier)
CONJ　等位接続詞 (coordinating conjunction)

タグ一覧：品詞タグ II

D　限定詞 (determiner)
FN　形式名詞 (formal noun)
FW　他言語の要素 (foreign word)
INTJ　間投詞 (interjection)
MD　モーダル要素 (modal element)
N　名詞 (noun)
NEG　否定辞 (negation)
NPR　固有名詞 (proper noun)
NUM　数詞 (numeral)
P　助詞 (particle)
P-COMP　補文助詞 (complementizer)
P-CONN　接続助詞 (conjunctional particle)
P-FINAL　終助詞 (final particle)
P-INTJ　間投助詞 (interjectional particle)

タグ一覧：品詞タグ III

P-OPTR　とりたて助詞 (toritate particle)
P-ROLE　格助詞 (role particle)
PASS　受動助動詞 (passive)
PNL　連体詞 (prenominal)
PRO　代名詞 (pronoun)
Q　量化詞 (quantifier)
SYM　記号 (symbol)
VB　動詞 (語幹) (verb (or verb stem))
VB0　軽動詞 (light verb)
VB2　補助動詞 (secondary verb)
WADV　疑問副詞 (indeterminate adverb)
WD　疑問限定詞 (indeterminate determiner)
WNUM　疑問数詞 (indeterminate numeral)
WPRO　疑問代名詞 (indeterminate pronoun)

タグ一覧：句タグ I

ADVP　副詞句 (adverb phrase)
CONJP　接続詞句 (conjunction phrase)
CP-EXL　感嘆節 (exclamative)
CP-FINAL　終助詞節 (projection for sentence final particle)
CP-QUE　疑問節 (直接または間接) (question (direct or indirect))
CP-QUE-ADV　副詞的な疑問節 (question used adverbially)
CP-QUE-OB1　目的語として用いられた疑問節 (question used as object)
CP-QUE-PRD　述語として用いられた疑問節 (question used as a nominal predicate)
CP-THT　補部節 (complementizer clause)
CP-THT-ADV　副詞的な THAT 節 (quote used adverbially)
CP-THT-SBJ　主語として用いられた THAT 節 (quote used as subject)
FRAG　断片 (fragment)

タグ一覧：句タグ II

FS　開始誤り (false start)
INTJP　間投詞句 (interjection phrase)
IP-ADV　副詞節 (adverbial clause)
IP-ADV2　主語により必ずコントロールされる副詞節 (obligatorily-subject-controled adverbial clause)
IP-ADV-CND　条件節 (conditional clause)
IP-ADV-CONJ　等位的な節 (coordinated clause)
IP-ADV-SCON　従属的な節 (subordinate clause)
IP-EMB　空所なし名詞修飾節 (gapless noun-modifying clause)
IP-IMP　命令節 (imperative clause)
IP-SMC　小節 (small clause)
IP-MAT　主節 (matrix clause)
IP-REL　関係節 (relative clause)
IP-SUB　準主節 (clause under CP* layer)

タグ一覧：句タグ III

IP-NML　名詞化節 (nominalized clause)
multi-sentence　多重文 (multiple sentence)
NML　中間名詞句 (intermediate nominal layer)
NP　名詞句 (noun phrase)
NP-ADV　副詞的名詞句 (adverbial noun phrase)
NP-LGS　論理的主語名詞句 (logical subject noun phrase)
NP-LOC　場所名詞句 (locational noun phrase)
NP-MSR　数量名詞句 (measure noun phrase)
NP-OB1　第一目的語名詞句 (primary object noun phrase)
NP-OB2　第二目的語名詞句 (second object noun phrase)
NP-POS　所有名詞句 (possessive noun phrase)
NP-PRD　述語名詞句 (predicate noun phrase)
NP-SBJ　主語名詞句 (subject noun phrase)
NP-SBJ2　第二主語名詞句 (secondary subject noun phrase)

タグ一覧：句タグ IV

NP-TMP　時間名詞句 (temporal noun phrase)
NP-TPC　主題名詞句 (topic noun phrase)
NP-VOC　呼格名詞句 (vocative noun phrase)
NUMCLP　助数詞句 (numeral-classifier phrase)
PNLP　連体句 (prenominal phrase)
PP　助詞句 (particle phrase)
PP-ADV　副詞的助詞句 (adverbial particle phrase)
PP-CMPL　補語的助詞句 (complement particle phrase)
PP-CND　条件を表す助詞句 (conditional particle phrase)
PP-CONJ　等位的助詞句 (coordination particle phrase)
PP-CZZ　被使役者助詞句 (causee particle phrase)
PP-LGS　論理的主語助詞句 (logical subject)
PP-LOC　場所助詞句 (locational particle phrase)
PP-MSR　数量助詞句 (measurement particle phrase)

タグ一覧：句タグ V

PP-OB1　第一目的語助詞句 (primary object particle phrase)
PP-OB2　第二目的語助詞句 (secondary object particle phrase)
PP-PRD　述語助詞句 (predicate particle phrase)
PP-PRP　目的助詞句 (purpositive particle phrase)
PP-SBJ　主語助詞句 (subject particle phrase)
PP-SBJ2　第二主語助詞句 (second subject particle phrase)
PP-SCON　従属節助詞句 (subordination particle phrase)
PP-TMP　時間助詞句 (temporal particle phrase)
PP-TPC　主題助詞句 (topical particle phrase)
PP-VOC　呼格助詞句 (vocative particle phrase)
PRN　括弧挿入句 (parenthetical)

タグ一覧：その他のタグ I

LS　リスト (list item)
LST　リスト項目 (list)
META　メタ情報 (meta information)
PU　句読点 (punctuation)

文字列検索

- 画面上部の 文字列検索 をクリック
- テキストの文字列を検索する
  - 文がどのように分節され，各々の要素にどのようなタグが与えられているか定かでない時は単純に文字列で検索することが有効

文字列検索

- 検索オプション：

	文字列の最初と最後は，語境界と一致しなくてもよい	文字列の最初と最後は，ある語の語頭と語末に一致する
文字列の中に語境界があるか否かを指定しない	Liberal	Character
文字列の中の語境界を半角スペースで指定する	Mine	Strict

- 画面下部：
  検索結果は CSV 形式でダウンロードが可能
- 各文のデータ番号をクリックすると，ツリーが表示される。

文字列検索

例：「ビル」という文字列を含んだ例を 4 つのオプションで検索してみてください。

例：「誰も」という文字列を Liberal で検索してみてください。

- 画面上，Tree fragments をクリックすると，文字列が作る部分的なツリーを表示する。

よくばり文字列検索

「誰も」を Liberal オプションで検索すると：

- 「誰も」の他に，「誰もが」のように助詞の「が」の付いた例がある（そういえば「誰とも」「誰からも」のように助詞が「誰」と「も」の間に現れることもある...）

- このように，指定した文字列の中に他の文字が入った例も検索したい場合は，よくばり文字列検索を使うとよい。

よくばり文字列検索

- 画面上部の よくばり文字列検索 をクリック
- 適切な検索オプションを選ぶ
  - 例えば「誰（...）も」を検索したい場合，「誰」はひとつの単語であり，「も」は一語，あるいは「でも」のような語の末尾であることが分かっているので，Character を選べばよい。
- 次に，割り込む文字数を選ぶ（デフォルトは 0）
  - Character オプションで，割り込む文字数を 2 にすれば，「誰」と「も」の間に 2 文字以下の文字が割り込んだ例を探す。
  - つまり，「誰も」以外に，「誰にも」「誰でも」「誰よりも」などが検索結果に含まれる
  - 割り込む文字数をさらに増やし，例えば，10 にすると，上記以外にも「誰が打っても」のような例もマッチする。
  - 割り込む文字数の制限はない。

文字列検索

- 文字列検索は，ある文字の連なりが NPCMJ の統語アノテーションに従って分類されて表示されるという点で便利
- 例えば，「という」は，どのようにアノテーションされているだろうか？
  - Liberal オプションで検索してみると，「という」は 15 のパターンでアノテーションがなされていることが分かる。

フレームグラフ

- 検索文字列のセグメンテーションと構造を視覚化したもの
  - 1 番下の空欄の行：マウスカーソルを合わせると、総トークン数が表示される。
  - 下から 2 番目の行：検索文字列のセグメンテーションのパターンが示される。マウスカーソルをブロックに合わせると、その実際の数が表示される。
  - 下から 3 番目より上の行：順により細かな下位分類が表示される。
    （フレームグラフの表示は Liberal、Character、Mine、Strict のどの検索オプションを選択したかによって変わる）