The Keyaki Treebank and the NPCMJ: Bridging a growing divide

Alastair Butler

NINJAL

The Keyaki Treebank has a simple format for encoding structural and semantic information with bracketed trees.

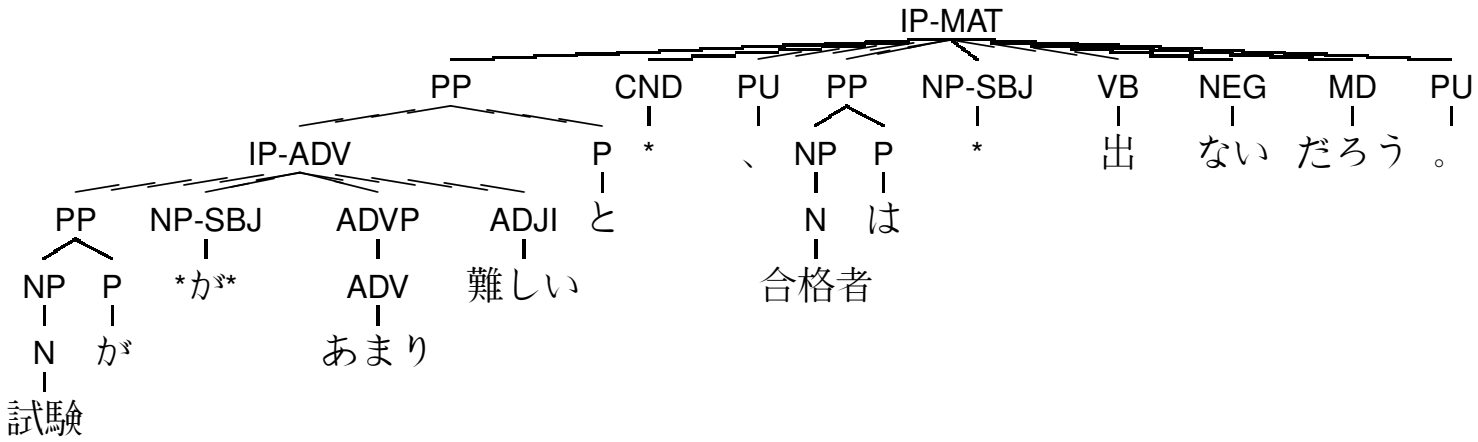
Aims to optimise tasks of annotation creation and maintenance.

The NINJAL Parsed Corpus of Modern Japanese (NPCMJ) has started as a repackaging of the Keyaki Treebank content into XML markup.

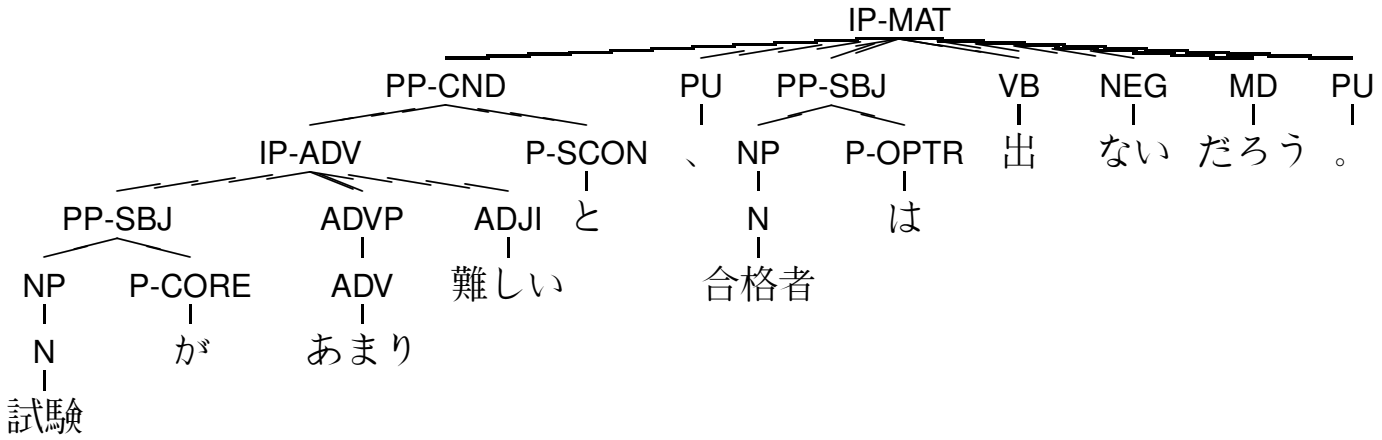
Aims to optimise tasks of search and presentation.

(reversible) alterations to how functional information is encoded

`parse_undecorate`

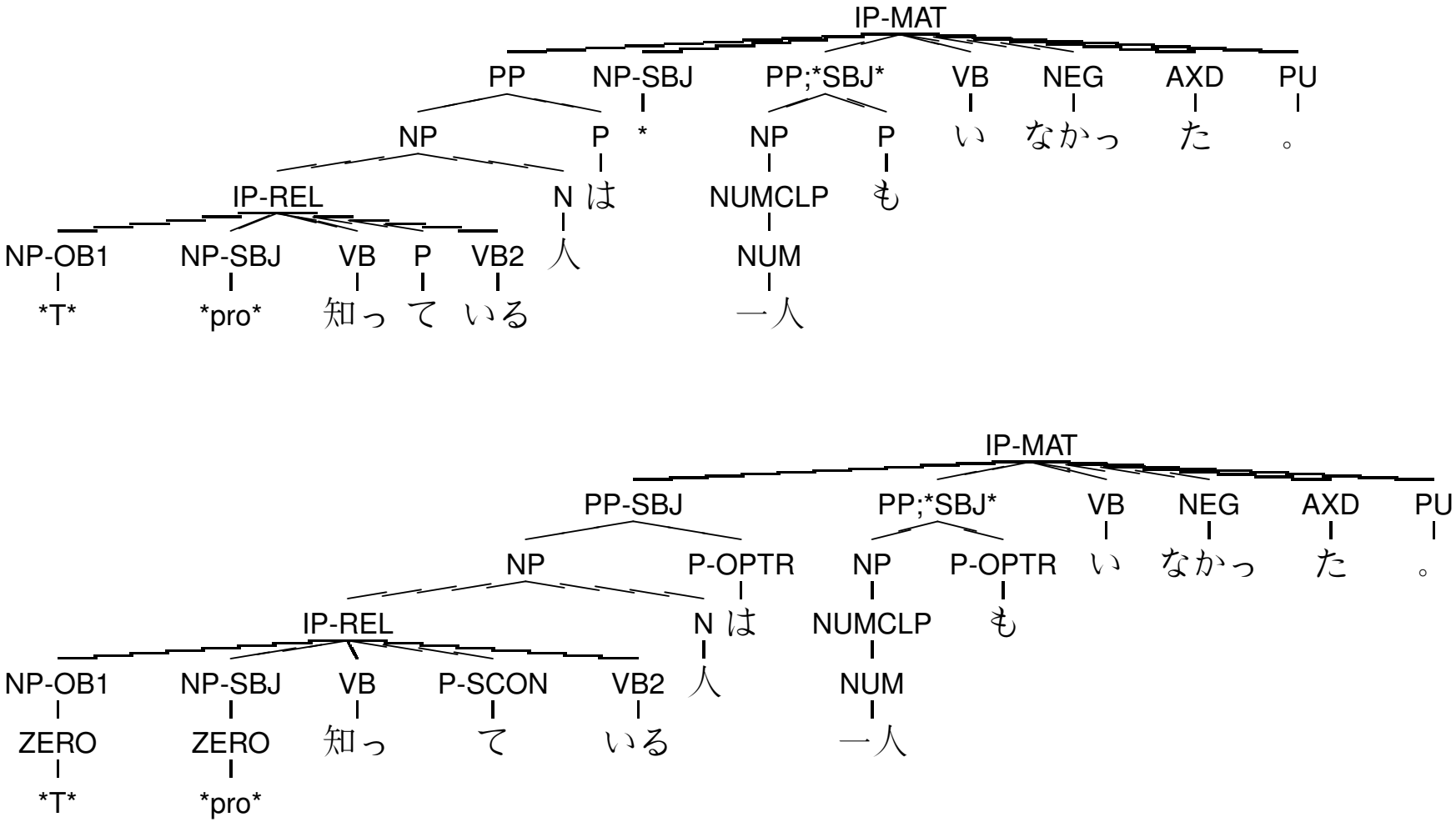


`parse_decorate`



<alpino_ds id="100_textbook_kisonihongo;page_43;MJ" version="1.3">
  <node cat="ip-mat" id="1" begin="0" end="12">
    <node cat="pp-cnd" id="2" begin="0" end="5">
      <node cat="ip-adv" id="3" begin="0" end="4">
        <node cat="pp-sbj" id="4" begin="0" end="2">
          <node cat="np" id="5" begin="0" end="1">
            <node pt="n" lemma="試験" aform="シケン" pos="名詞-普通名詞-サ変可能" pron="シケン" word="試験" id="6" begin="0" end="1"/>
          </node>
          <node pt="p-core" lemma="が" aform="ガ" pos="助詞-格助詞" pron="ガ" word="が" id="7" begin="1" end="2"/>
        </node>
        <node cat="advp" id="8" begin="2" end="3">
          <node pt="adv" lemma="余り" aform="アマリ" pos="副詞" pron="アマリ" word="あまり" id="9" begin="2" end="3"/>
        </node>
        <node pt="adji" lemma="難しい" aform="ムズカシイ" pos="形容詞-一般" pron="ムズカシー" stype="形容詞" vform="終止形-一般" word="難しい" id="10" b
end="4"/>
      </node>
      <node pt="p-scon" lemma="と" aform="ト" pos="助詞-接続助詞" pron="ト" word="と" id="11" begin="4" end="5"/>
    </node>
    <node pt="pu" lemma="、" pos="補助記号-読点" word="、" id="12" begin="5" end="6"/>
    <node cat="pp-sbj" id="13" begin="6" end="8">
      <node cat="np" id="14" begin="6" end="7">
        <node pt="n" lemma="合格 者" aform="ゴウカク シャ" pos="名詞-普通名詞-サ変可能 接尾辞-名詞的-一般" pron="ゴーカク シャ" word="合格者" id="15" begin="6" en
      </node>
      <node pt="p-optr" lemma="は" aform="ハ" pos="助詞-係助詞" pron="ワ" word="は" id="16" begin="7" end="8"/>
    </node>
    <node pt="vb" lemma="出る" aform="デル" pos="動詞-一般" pron="デ" stype="下一段-ダ行" vform="未然形-一般" word="出" id="17" begin="8" end="9"/>
    <node pt="neg" lemma="ない" aform="ナイ" pos="助動詞" pron="ナイ" stype="助動詞-ナイ" vform="終止形-一般" word="ない" id="18" begin="9" end="10"/>
    <node pt="md" lemma="だ" aform="ダ" pos="助動詞" pron="ダロー" stype="助動詞-ダ" vform="意志推量形" word="だろう" id="19" begin="10" end="11"/>
    <node pt="pu" lemma="。" pos="補助記号-句点" word="。" id="20" begin="11" end="12"/>
  </node>
  <sentence>試験 が あまり 難しい と 、 合格者 は 出 ない だろう 。</sentence>
</alpino_ds>

Zero elements get a POS node



```
<alpino_ds id="738_textbook_kisonihongo;page_142;MJ" version="1.3">
  <node cat="ip-mat" id="1" begin="0" end="13">
    <node cat="pp-sbj" id="2" begin="0" end="7">
      <node cat="np" id="3" begin="0" end="6">
        <node cat="ip-rel" id="4" begin="0" end="5">
          <node cat="np-ob1" id="5" begin="0" end="1">
            <node pt="zero" word="*T*" id="6" begin="0" end="1"/>
          </node>
          <node cat="np-sbj" id="7" begin="1" end="2">
            <node pt="zero" word="*pro*" id="8" begin="1" end="2"/>
          </node>
          <node pt="vb" word="知っ" id="9" begin="2" end="3"/>
          <node pt="p-scon" word="て" id="10" begin="3" end="4"/>
          <node pt="vb2" word="いる" id="11" begin="4" end="5"/>
        </node>
        <node pt="n" word="人" id="12" begin="5" end="6"/>
      </node>
      <node pt="p-optr" word="は" id="13" begin="6" end="7"/>
    </node>
    <node cat="pp;*sbj*" id="14" begin="7" end="9">
      <node cat="np" id="15" begin="7" end="8">
        <node cat="numclp" id="16" begin="7" end="8">
          <node pt="num" word="一人" id="17" begin="7" end="8"/>
        </node>
      </node>
      <node pt="p-optr" word="も" id="18" begin="8" end="9"/>
    </node>
    <node pt="vb" word="い" id="19" begin="9" end="10"/>
    <node pt="neg" word="なかっ" id="20" begin="10" end="11"/>
    <node pt="axd" word="た" id="21" begin="11" end="12"/>
    <node pt="pu" word="。" id="22" begin="12" end="13"/>
  </node>
  <sentence>*T* *pro* 知っ て いる 人 は 一人 も い なかっ た 。</sentence>
</alpino_ds>
```

While bracketed notation is limited to coding tree nodes as single character strings, XML is essentially limitless in its extendibility without needing to overload accessibility to node information.

This is proving to be extremely beneficial as the NPCMJ grows in the richness of the information offered.

The Keyaki Treebank itself can continue to be enlarged and refined in its native bracketed format and yet preserve its role as the pivotal component of a widening NPCMJ enterprise.