Parsed corpus annotation (ad)ventures

Alastair Butler, Iku Nagasaki, Stephen Wright Horn, Susanne Miyata, Zhou Zhen, Kei Yoshimoto

NINJAL, Aichi Shukotoku University and Tohoku University

Kobe Meeting, Nov 4, 2017

This talk is about parsed (treebank) annotation used for:

- Contemporary Japanese (Keyaki Treebank, NPCMJ)
- English (TSPC)
- Old Japanese (M97PC)
- Kolyma Yukaghir
- Child Japanese
- Contemporary Chinese
- ...

Talk overview

- aims of building parsed annotation
- background
- the contribution of accompanying interpretation
- annotation for Japanese
- how we start annotation

Our parsed annotation approach aims

- to provide fully searchable representations of texts that are described according to interpretations of the meanings of sentences.

This involves:

- analysing texts into units (NP-SBJ, PP-TMP, ADVP-MNR, IP-REL, etc.), and
- accounting for how the units are composed into meaningful expressions (with structural relations, and only as a last resort, indices),

On the shoulders of giants

- Penn Historical Corpus (Santorini 2010): ideas about phrase structure (a flattened X-bar), what nodes might be called, function marking, ideas about how to undertake annotation (scaffolding)
- SUSANNE Corpus and Lancaster approach (Sampson 1995): ideas about phrase structure (clause coordination represented as tag distinguished subordination), rich functional marking, rich morphological marking
- Alpino Corpus (van Noord 2002): XML encoding, internet based search (PaQu)

What's new?

- analysis (grammatical description) for different languages
- binding information (to resolve anaphoric dependencies)
- an interpretation procedure to ground the annotation
- almost entirely removes dependence on indexing
- an opening to integrate lexical semantics (PropBank/FrameNet)

Parsed annotation accompanied by interpretation

Parsed annotation becomes exponentially more revealing – and interesting – if accompanied by an interpretation mechanism that breathes life into the annotation.

With interpretation, a fuller description of a sentence can be achieved, by making aspects explicit that the raw annotation has left either implicit or underspecified.

A key objective of our annotation has been to reach a **critical mass** of organised information to support automatic interpretation procedures.

Automatic interpretation procedures can:

- from parsed annotation, reach a predicate logic / DRS representation
- from predicate logic / DRS representations, place indexing (bound variable) information into parsed annotation
- from indexed parsed annotation, obtain rich word dependency information
- from rich word dependency information, ...

Our annotation for Japanese

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

Aims to optimise tasks of annotation creation.

Has released 38,911 trees (24,478 with full data) for download.

Offers a detailed manual. Has BCCWJ style / UniDic morphological analysis.

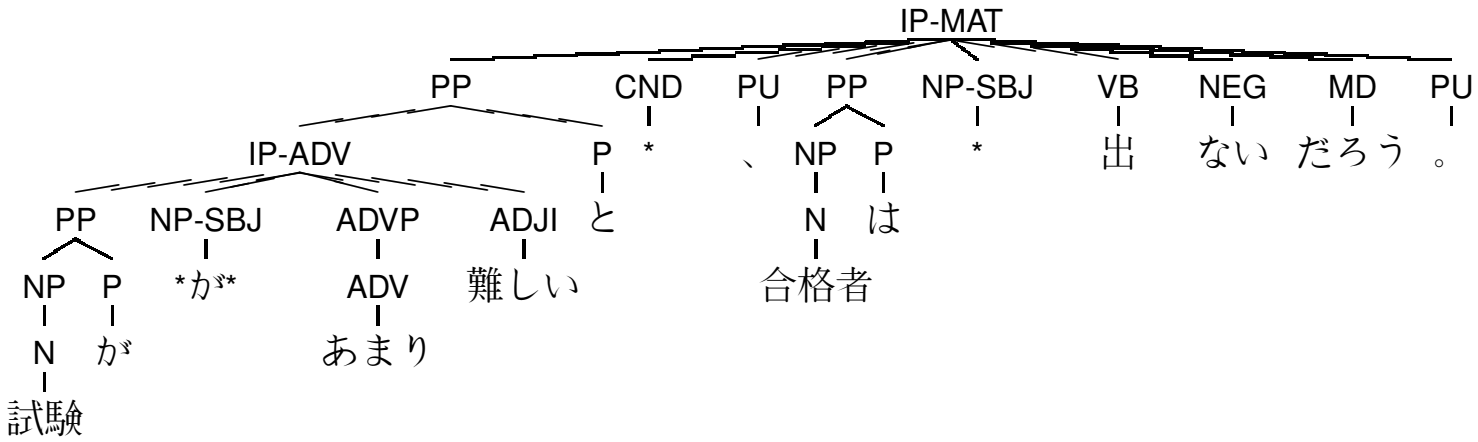
The **NINJAL Parsed Corpus of Modern Japanese** (NPCMJ) started as a repackaging and reworking of the Keyaki Treebank content.

Aims to optimise tasks of search and presentation.

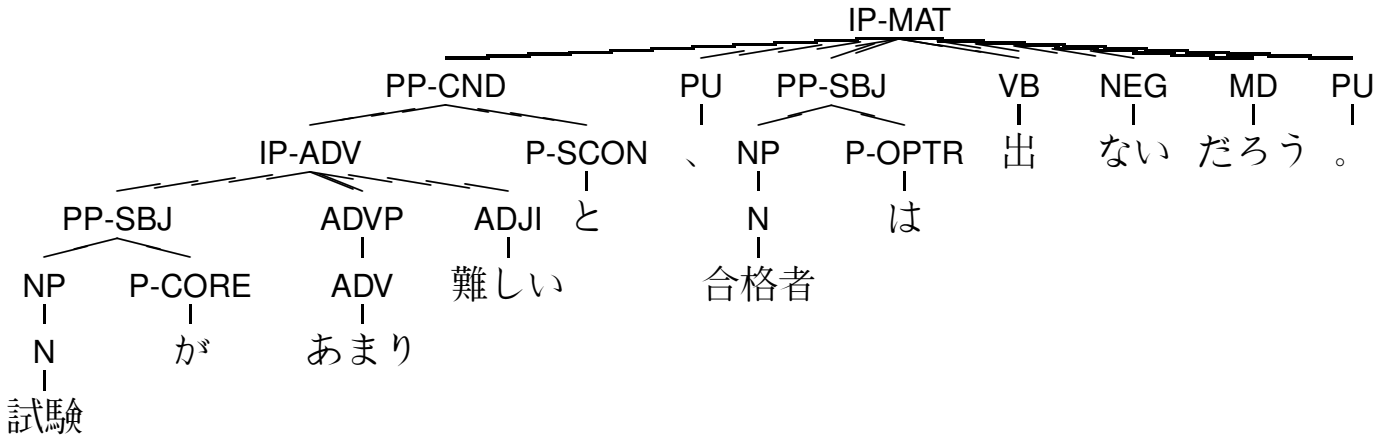
Has released 10,098 trees through a web interface.

(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>
```

Starting parsed annotation

- First, morphological analysis is required, to give segmented words and parts-of-speech
- Second, there needs to be projection of tree structure
  - from a parser trained on already existing parsed sentences
  - via a scaffolding script
- Third, there needs to be human correction and elaboration