

AI を用いた 中国語研究の可能性

『言語研究と生成系 AI』

日本中国語学会第 73 回全国大会, 2023/11/04

山崎直樹

(関西大学外国語学部／外国語教育学研究科)

概略

1. 探索的な使いかたいろいろ（考えるための手がかり）
2. 知識の必要な機械的作業の下請けとして使う
3. 独自規格の可能性を探る

主に使用した対話生成系AI

1. ChatGPT4（これに、いろいろなプラグインを組み込んでいます。有料版です）
2. Claude 2（無料版）

探索的な……

- 「膨大なデータをもとに、いちばん確率が高そうなことを、べらべらしゃべる」のが生成系AI
- 条件をあれこれ変えてどんなことをしゃべるのか、見てみよう

類義表現のヒント

(1) 次のa, b, cに後続する文をそれぞれ5文ずつ作ってください。

- a. 他是一個很善良的人。
- b. 彼は善良な人です。
- c. 彼はやさしい人です。

ある映画の日本語の「やさしい」が中国語の“善良”で置き換えられていたそうです……

(2) 次のa', b', c'に後続する文をそれぞれ5文ずつ作ってください。

- a'. 她是一個很善良的人。
- b'. 彼女は善良な人です。
- c'. 彼女はやさしい人です。

注意！

- コーパスと比べて、条件が統制しやすいけれど……
- 生成された表現のauthenticityについてはチェックが必要
- 検証が必要なのは、均衡性が怪しいコーパスでも同じですけどね

文脈と文型

ある特定の文型の文の前後にはどのような文が現れるかを知りたいです。
次の(a)(b)それぞれの文の前と後に現れる典型的な文を教えてください。

(a) 他學韓語學了三年。

(a)の**前**に現れる文:

(a)の**後**に現れる文:

(b) 他學了三年韓語。

(b)の**前**に現れる文:

(b)の**後**に現れる文:

方略的な談話の生成

- 映画に誘う
- レポートの締め切りを延ばしてもらう

映画に誘う

- わたしは日本人の女性です。
- 映画『咖啡時光』のチケットを2枚持っています。
- 次の土曜の午後、台湾人留学生の友人・魏吟玲さん（女性）を誘って2人でこの映画を見にいきたいと思います。
- 魏吟玲さんとは、世間話くらいはしますが、今までどこかにいっしょに出かけたことはありません。

- 陳怡靜（女性）を誘って……
- 陳怡靜とは、同じ年で仲が良く、今までもいろいろなところにいっしょに遊びに行っています

レポートの締め切りを延ばしてもらおう

- あなたは、インストラクショナル・デザイン（教學設計）の授業で出されたレポートの締め切りを、うっかり忘れていました。
- レポートのテーマは、「Project-based learning（專題導向學習）の長所と短所」です。
- そこで、授業の担当教師である王先生にレポートの締め切りを3日延ばしてもらおうお願いをするメールを書こうと思いました。
- 嘘でもよいので、何か理由を探して、お願いを試みようと思っています。
- このメールを台湾華語で書いてください。

余談：AIにも性格がある？

- 「不誠実な理由をつけることは、一般的に推奨される行動ではありません。しかし、あくまでシチュエーションを想定してのリクエストと捉え……現実の状況で不正直な理由を使って延期を求める行動は、信頼を損なう恐れがありますので注意が必要です」 (ChatGPT4)
- Claude2のほうが、一般に、（頼んでいないことまで）ていねいに説明してくれる傾向がある……が、軽薄な印象がぬぐえない（お世辞も言うし……）。「人間性」に問題がある？

知識の必要な機械的作業の下請け

- 正規表現
- Pythonによるテキスト解析
- 文構造の解析とアノテーション
- テキスト構造の解析とアノテーション

正規表現

次のパターン全てにヒットする正規表現を書いてもらう

- 可不可以，喜不喜歡，漂不漂亮，學沒學習……
- `(.)(不)\1.`
- これだと、“好不好”などにもヒットしますがどね

テキスト解析

- ◆ n-gram (連続する任意のn文字で頻出するパターン)
- ◆ skip-gram (特定のテキストにおいて、そのキーワードの前後それぞれn語には、どのような表現が頻出するか)
- ◆ ChatGPTが処理を実行してくれる。望めば、処理に使えるPythonのスク립トを見せてくれる。さらに望めば、スク립トの部分部分に関する質問に答えてくれる。もしスク립トを実行してエラーが出たら、そのエラーメッセージを解説してくれる

注意……

スクリプトがまちがっていても、誰も責任を取ってくれないので、自分でそのスクリプトを検証しておく必要がある

「自分ではぜんぜんできないことを代行させる」こともできるが、その結果に責任を負うのは自分なので、「自分で検証できること」の範囲内で下請けに出すべき

文構造の解析とアノテーション

- 品詞や構成素構造に関するメタデータを、テキストの中にタグとして埋め込んだコーパス、あるいは、テキストを解析してそのようなメタデータを表示してくれるシステムは、いろいろありますが……
- 今どきのAIは、有名なものについては、みなその仕様を知っている……つまり、代わりにタグ付けをしてくれる（**まちがえることも多々あります！**）

例えば……

- 中央研究院現代漢語標記語料庫
- 北京大学现代汉语语料库
- Penn Chinese Treebank

中央研究院現代漢語標記語料庫

你千萬不要把脫下來的襪子放在餐桌上

你 (nǐ) - Pronoun (PN)

千萬 (qiān wàn) - Adverb (AD)

不要 (bù yào) - Modal verb (MV)

把 (bǎ) - Preposition (P)

脫下來 (tuō xià lái) - Verb phrase (VP)

的 (de) - Particle (DE)

襪子 (wà zi) - Noun (N)

放 (fàng) - Verb (V)

在 (zài) - Preposition (P)

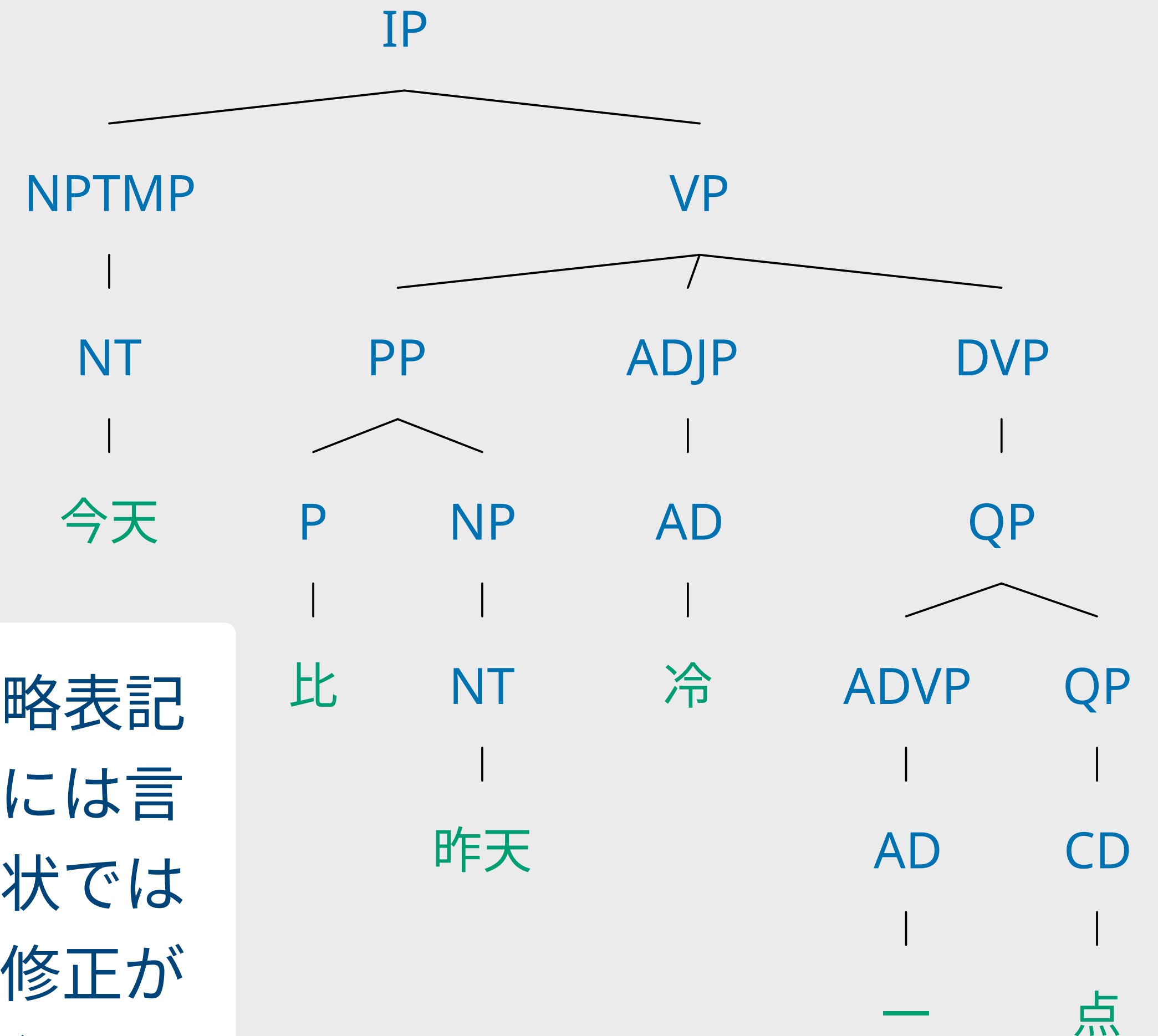
餐桌 (cān zhuō) - Noun (N)

上 (shàng) - Noun (N) or Postposition (POST)

精密表記と簡略表記があり、一概には言えないが、現状ではガイドラインどおりにならないことも多く、手作業での修正が（かなり）必要

Chinese Treebank

(IP
(NP-TMP
(NT 今天))
(VP
(PP
(P 比)
(NP
(NT 昨天)))
(ADJP
(AD 冷))
(DVP
(QP
(ADVP
(AD 一))
(QP
(CD 点))))))



精密表記と簡略表記
があり、一概には言
えないが、現状では
手作業による修正が
(かなり) 必要

Chinese Propbank

你千萬不要把
脫下來的襪子
放在餐桌上

この文の述語は「放/fang」(置く)で、3つのコア引数を取ります

Arg0: 置く主体 (agent)

Arg1: 置かれるもの (theme)

Arg2: 置かれる場所 (location)

Arg0: 省略 (you)

Arg1: 脱いだ靴下 (脱下來的襪子)

Arg2: 食卓の上 (在餐桌上)

と解析できます。アノテーションは以下のようになります:

[Arg0 你] 千萬不要 [Arg1 把脫下來的襪子] [Rel 放] [Arg2 在餐桌上]

注意！

- 「学習なし」でもそれらしい出力をしてくれるが、怪しい
- タグ付けのガイドラインが、読み取り可能なPDFになっていれば、それを読ませたほうがよい（ChatGPT4はプラグインでその機能を提供、Claude2は、デフォルトでその機能をもっている）
- うまく読み取れないことも多い

何がうれしいか

- このようなタグ付けが半自動化されれば、自分だけのコーパスができるかも。
- PDF文書の読み取りに問題があることも多いと思われるが、早晩、改善されると思う……

テキスト構造のアノテーション(1)

- Text Encoding Initiative (<https://tei-c.org/>) による文献の電子的表現と交換のためのアノテーションガイドライン

〈原文〉

1) 國破山河在 2) 城春草木深 3) 感時花
濺淚 4) 恨別鳥驚心 5) 烽火連三月 6)
家書抵萬金 7) 白頭搔更短 8) 渾欲不勝簪

〈メタデータ〉

「五言絶句」。詩の題は「春望」。唐の時代の詩。作者は「杜甫」。1)と2)は対句です。3)と4)は対句です。5)と6)は対句です。脚韻を踏んでいるのは「深, 心, 金, 簪」です。

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>春望</title><author>杜甫</author>
      </titleStmt>
      <publicationStmt>
      </publicationStmt>
      <sourceDesc>
        <p>唐代の五言律詩</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <lg type="五言律詩">
        <lg type="couplet">
          <l>國破山河在</l><l>城春草木<rhyme>深</rhyme></l>
        </lg>
        <lg type="couplet">
          <l>感時花濺淚</l><l>恨別鳥驚<rhyme>心</rhyme></l>
        </lg>
        <lg type="couplet">
          <l>烽火連三月</l><l>家書抵萬<rhyme>金</rhyme></l>
        </lg>
        <l>白頭搔更短</l><l>渾欲不勝<rhyme>簪</rhyme></l>
      </lg>
    </body>
  </text>
</TEI>
```


是為太祖今王是太祖曾孫亦生視
魏書云朱蒙裔孫宮生而開目能視
山上王諱延優一名位宮故国川王之弟也

[本文]に示したテキスト
に続いて、[割注]に示し
た注が続きます。これ
をTEIのガイドラインに
従って、タグ付けをし
てください。

[本文]山上王諱延優

[割注]一名位宮

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
  </teiHeader>
  <text>
    <body>
      <div type="historicalRecord">
        <p>
          <app>
            <lem>山上王諱延優</lem>
            <rdg>一名位宮</rdg>
          </app>
        </p>
      </div>
    </body>
  </text>
</TEI>
```

テキスト構造のアノテーション(2)

- Rhetorical structure theory (RST)
- 試しに、Wikipediaにある“律詩”の項の冒頭を分析してみると……
(ChatGPT4)

1. 南北朝文學作品注重對偶和平仄，對律詩的發展影響深遠。

Background

└─ 2. 南齊永明之世，周顥和沈約創四聲八病之說。

Elaboration

Result

└─ 3. 影響所及，當時詩人作詩講究聲律，世稱「永明體」，這是由古體詩……

└─ 4. 唐代以前的古體（或稱古風）律詩，其長度不限。

Background

Elaboration

└─ 5. 唐初時期，上官儀提倡「六對」、「八對」之說……對律詩的發展頗有影響。

└─ 6. 初唐四傑所作律詩甚多，促進了律詩的發展。

Cause

└─ 7. 沈佺期、宋之問在前人實踐的基礎上，加以琢磨、去粗取精……

└─ 8. 杜甫是第一位大量創作七言律詩詩人。

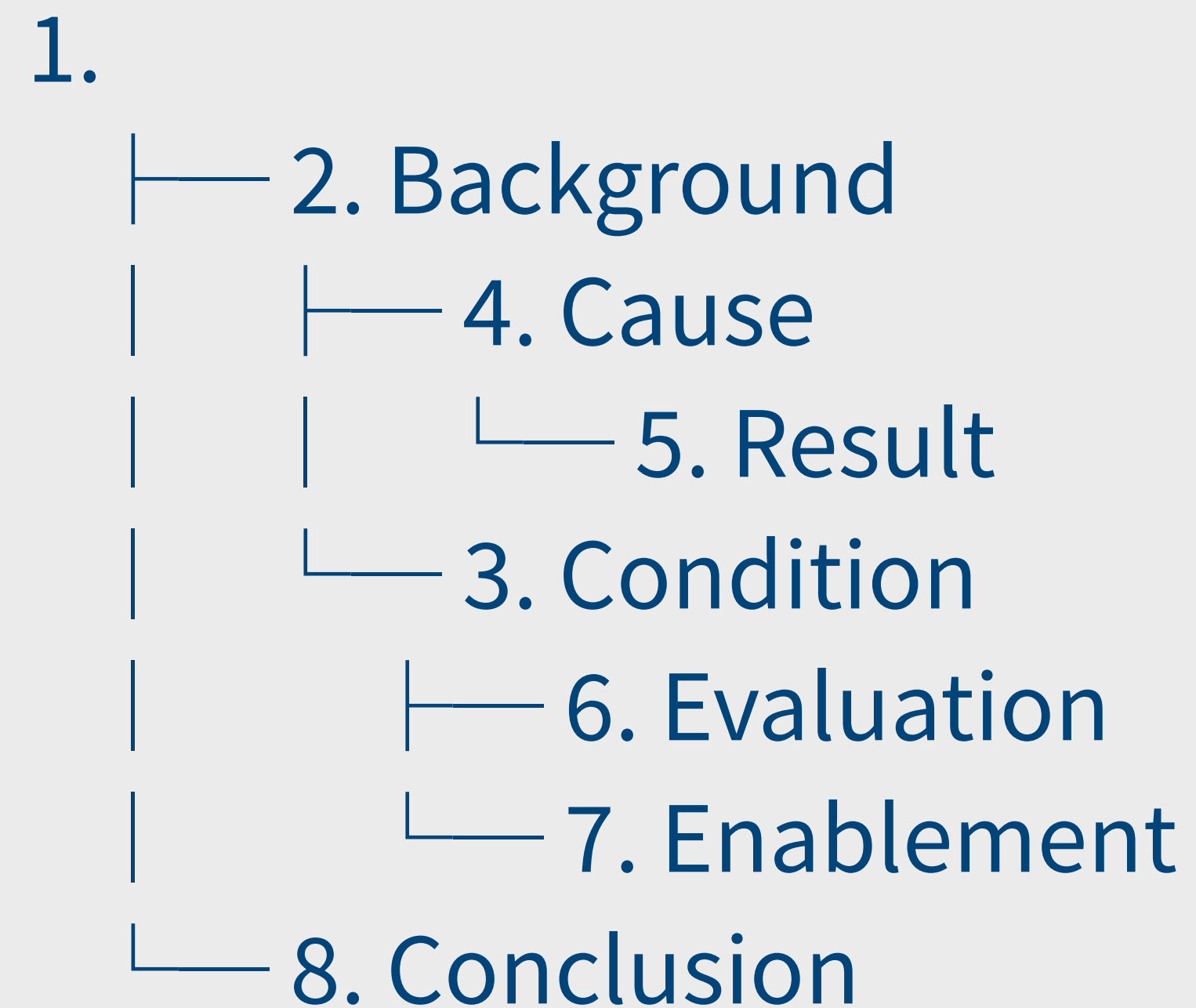
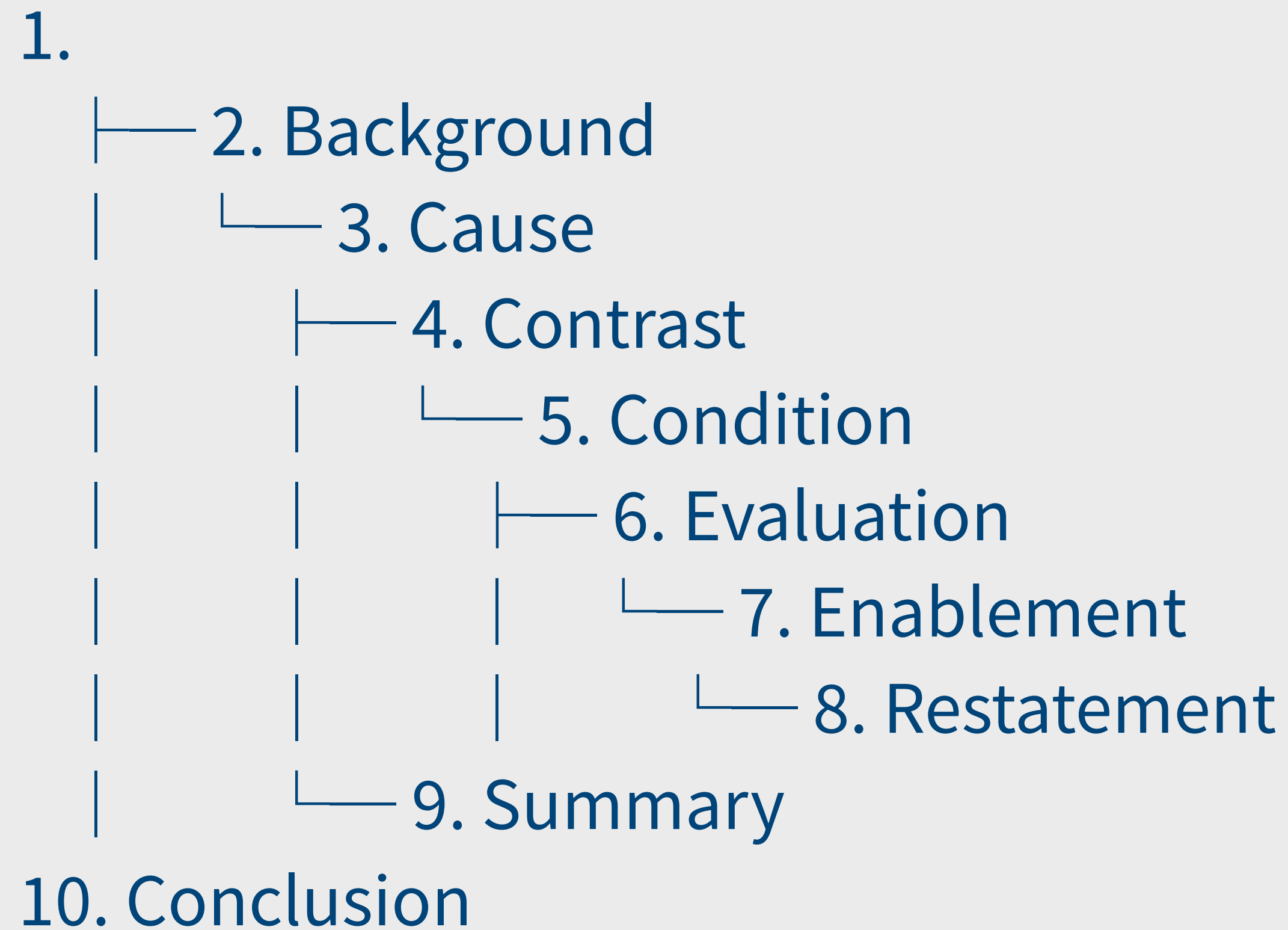
Summary

Elaboration

レポートの締め切りを延ばしてもらおう

- あなたは、インストラクショナル・デザイン（教學設計）の授業で出されたレポートの締め切りを、うっかり忘れていました。
- レポートのテーマは、「Project-based learning（專題導向學習）の長所と短所」です。
- そこで、授業の担当教師である王先生にレポートの締め切りを3日延ばしてもらおうお願いをするメールを書こうと思いました。
- 嘘でもよいので、何か理由を探して、お願いを試してみようと思っています。

複数のテキストの構造比較



母語話者の意見(1)

〈出力1〉は「謝罪」 (“虽然这是我的责任……学习成果。”) のタイミングが早すぎるし、長い感じがしました (このタイミングで謝るのなら“对不起/实在抱歉”ぐらいに簡単な言葉にしたほうが)。「提案」 (“延长报告的截止日期3天”) までの文章が長すぎたので、何をしてほしいのかまでをもう少し短くしてほしいと私は思いました。文章の最後のほうも必要以上に謝っているので、近い過去にも宿題を提出し忘れて、同じような過ちを繰り返している人なのかなと思ってしまいました。

母語話者の意見(2)

两个相比，我认为〈出力1〉更合适。理由如下：1) 首先〈出力1〉进行了自我介绍，而〈出力2〉没有，不给老师做自我介绍有一点没有礼貌（虽然“亲爱的王老师”有点奇怪。）2) 然后，〈出力1〉先给老师道歉，然后陈述了客观和主观的理由，最后提出了自己的请求，这样子的顺序读者很容易共情；〈出力2〉虽然陈述得很明确道歉也很诚恳，但没有给出足够的理由就提出“我想请您延后三天的时间”，听起来有点任性；3) 作为道歉的理由，我绝对不会使用〈出力2〉，诅咒自己家人生病我觉得很不好。〈出力1〉“家中突发的事情”，我相比更容易接受，4) 文章最后，〈出力2〉用了“祝安好”，这个对老师用有点没有礼貌。

方略的談話の要点

- どんな内容を
- どんなタイミングで
- どんな量で

余談：2言語間の距離

- Glossingの使い途
- 文単位のデータを示すとき、「原文—単語単位での意味—全体の意味」を並べて示す
- Leipzig glossing rules

余談：2言語間の編集距離

1) 香香 は 蕾蕾 より 4 歳 年上 だ

Xiangxiang TOP Leilei COMP 4 age older COP

“Xiangxiang is 4 years older than Leilei.”

gloss単位で編集距離を取れば、2種の構造間の距離が測れる？

2) 香香 比 蕾蕾 大 4 歳

Xiangxiang COMP Leilei big 4 age

“Xiangxiang is 4 years older than Leilei.”

glossingも距離の計算もAIがやってくれる……

独自規格の可能性を探る

- いくつかの生成系AIは、ウェブ上のデータや特定のファイル形式の文書を読み取り、その内容を要約したり、その内容に沿って指示に答えてくれます
- 自分独自の規格を作って、AIに学ばせたら……

余談：文法情報をもつコーパスの設計

- 品詞情報くらいならまだよいのですが、統語的解析を必要とする情報をもたせようとする、その仕様をめぐって争いが生じます。
- 解決策(1): ひとりでやる
- 解決策(2): 言語研究者を仲間に入れない

独自の品詞体系

昔、古典中国語の機械解析のために、こんな品詞分類表を作りました
(国語研の分類語彙体系を参考にしました。ある種のfacet分類ですね)

魯: 名詞-主体-国名

山: 名詞-固定物-地形

重: 動詞-描写-量

往: 動詞-行為-移動

.....

これを学習させたら……

孟子: 名詞 人 複合的人名

見: 動詞 行為 視覺

梁: 名詞 主体 国名

惠: 名詞 人 その他の人名

王: 名詞 人 役割

王: 名詞 人 役割

曰: 動詞 行為 伝達

叟: 名詞 人 人

不: 副詞 否定 無界

遠: 動詞 描写 量

千: 数詞 数字

里: 名詞 度量衡

而: 助詞 接続 並列

來: 動詞 行為 移動

亦: 副詞 否定 無界

將: 副詞 時相 将来

有: 動詞 存在 存在

以: 前置詞 源泉

利: 動詞 行為 得失

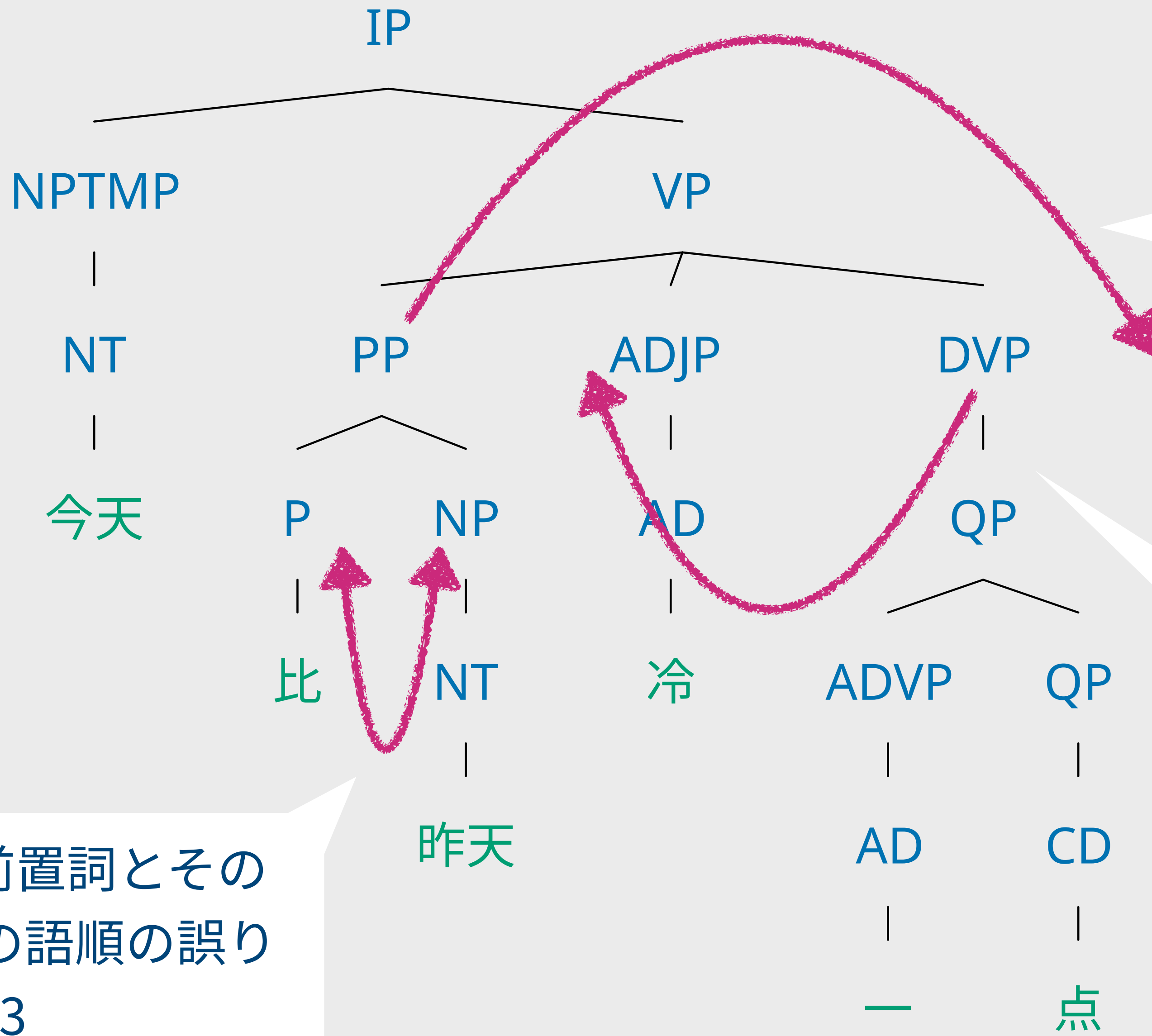
吾: 代名詞 人称 起格

國: 名詞 主体 集團

乎: 助詞 句末

こんな実験をしました

1. 文構造の誤用の判定基準ガイドラインを作る（もちろん独自規格）
2. 生成系AIにそれを学ばせる
3. 誤用例を与えて、判定させる



TYPE: 前置詞句の位置の誤り
深刻度: 2

TYPE: 数量詞句の位置の誤り
深刻度: 2

TYPE: 前置詞とその目的語の語順の誤り
深刻度: 3

実験

1. 学習用データ: 今天比昨天冷一點 / *今天冷一點比昨天 / *今天比昨天一點冷 / *今天昨天比冷一點
2. なぜ誤用かの説明: (PP)が(DVP)の後にある / (QP)が(ADJP)の前にある / (PP)の下の(P)がNPの後にある
3. 評価用データ: *哥哥比我两岁大 / *這個比那個10元貴 / *我妹妹，我比高10釐米 / *我重很多比老婆

結果と考察

1. 1回ではうまくいかない。修正フィードバックが必要（学習用データが1例で、評価用は初見なのでやむをえない）
2. フィードバックの後、語彙が異なる同じタイプの誤用例を与えたら、1回で正解に達した
3. 誤用の描写は、わざと「前にある／後にある」のように定義したが、形式的に表現すればもっと理解しやすいかも。

結び

- 自分で作った規格は、完全に「自分でチェックできる」範囲内なので、歴史に名を残したい人は、ぜひ……
- 普及する／しないのカギは、中身よりも、順番（最初に出たものが普及することが多い）
- 今日の話で言及した規格・フレームワーク・理論については → <https://scrapbox.io/2023-11-04/>

[https://scrapbox.io/](https://scrapbox.io/2023-11-04/)



2023-11-04/