レトリック言語資源整備の戦略

九州大学　言語文化研究院
伊藤　薫

2024/3/15

日本語言語資源の構築と利用性の向上 (JLR2024)

はじめに

自己紹介・研究背景

- テーマ：多様な言語資源の領域間融通
  - Universal Dependenciesとの関わり
- 発表者はレトリック研究が専門
  - 以前は認知意味論ベースの分析をやっていました
  - 現在は構文が関わる修辞表現の言語資源を構築中
- 本発表の**目的**
  - レトリックに関わる言語資源を取り巻く環境を確認し、これからのあり方を議論
  - コーパス構築を済ませ、それを紹介するわけではない

言語資源を巡る動向

- LLMの時代
- NLPではAIが基本的な文法を「マスター」し、自然言語理解や実世界接地など、高度な言語能力へ注目が集まる
- 課題に取り組むためには言語資源が不可欠

レトリックとは

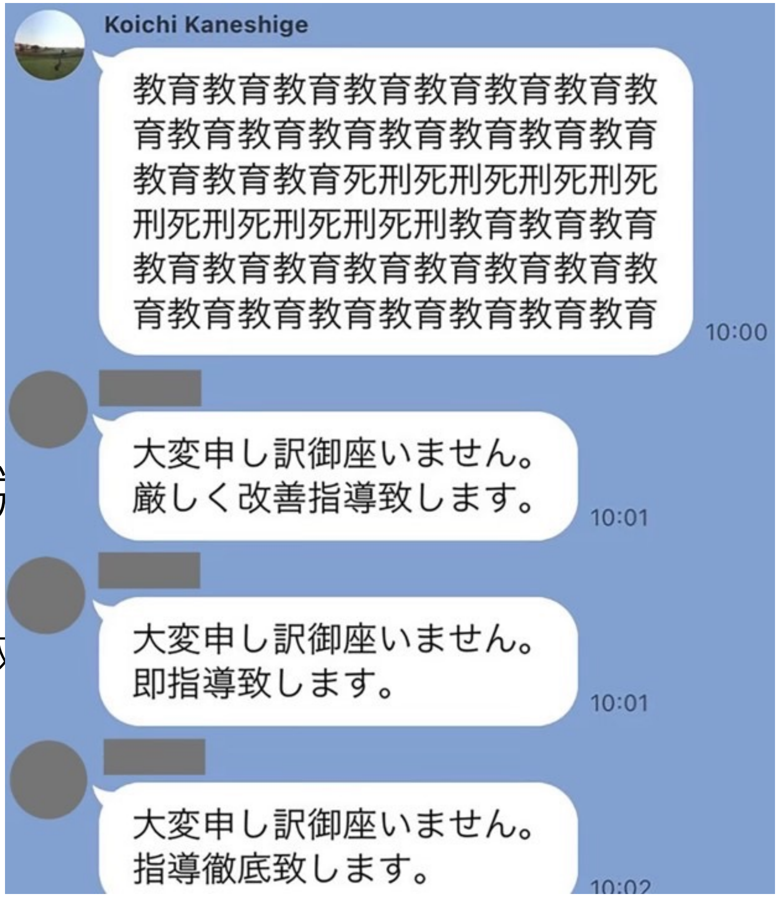
- 弁論術と修辞学
  - 上手く話すための技術
  - 本発表では修辞に絞って述べる
- 伝統的な修辞学では様々な効果的な表現の「型」を整理
  - 一般的な表現からの逸脱のパターン [佐藤92]
- 近年の言語学では身体性や認知能力との関わりで研究 [Lakoff+80]
- 具体例
  - メタファー「SNSの発言が**炎上する**」
  - メトニミー「**鍋**を食べる」「**ブラームス**を聴く」
  - 平行法 “I am fond of pigs. **Dogs look up to us. Cats look down on us.** Pigs treat us as equals.” (Winston Churchill)
  - 緩叙法「この味、**嫌いじゃない**よ」
  - 異義兼用「あそこの店、**影も味も薄い**な」
  - 誇張法「**100年ぶり**にコーヒー飲んだ」

レトリカルな生成AI

- 動機：同じ内容を伝えるのであれば・・・
  - 分かりやすい方がいい・面白い方がいい・説得力のある方がいい
  - 例
    - 運営業務と教育業務が多すぎて研究する時間が取れない。（普通の文）
    - 山のように他の仕事が降ってきて研究できない。（直喩と隠喩）
    - 運営、雑用、メール、問い合わせ、授業、採点、シラバス作成、会議。一体いつ研究すればいいんだ。（列叙法、反語法）
    - 教育教育教育教育教育教育雑用雑用雑用研究雑用雑用雑用メールメールメールメールメール（反復法、暗示引用）
- 効果的な文章の生成と評価、分析にレトリック言語資源が必要

(https://friday.kodansha.co.jp/article/324316/photo/c05195c1)

レトリカルな生成AI



- 動機：同じ内容を伝えるのであれば・・
  - 分かりやすい方がいい・面白い方がいい・
  - 例
    - 運営業務と教育業務が多すぎて研究する時間が取
    - 山のように他の仕事が降ってきて研究できない。
    - 運営、雑用、メール、問い合わせ、授業、採点、研究すればいいんだ。（列叙法、反語法）
    - 教育教育教育教育教育教育雑用雑用雑用研究雑用雑用雑用メールメールメールメールメール（反復法、暗示引用）
- 効果的な文章の生成と評価、分析にレトリック言語資源が必要

修辞技法の性質

- 定義が難しく、データも作りにくい
  - 機械的に探しづらく、人によってもばらつきが出る
  - そもそも専門書間でも微妙に定義や分類が異なる
  - e.g. 創造的なメタファーと死喩の区別
  - 「元ネタ」を知らないと気付けない修辞表現も (e.g. 暗示引用)
- マイナーな修辞技法は認知度が低い
  - ただし、知らないうちに日常で接してはいる

レトリック言語資源の利用者

言語資源の利用者

- NLP研究者
  - 機械学習（学習・評価 / 検証）→言語モデル、サービス
- 言語学者
  - 量的研究
  - 質的研究
- デジタル・ヒューマニティーズ
  - 資料としての価値
  - 人が閲覧

言語学での構築・利用状況

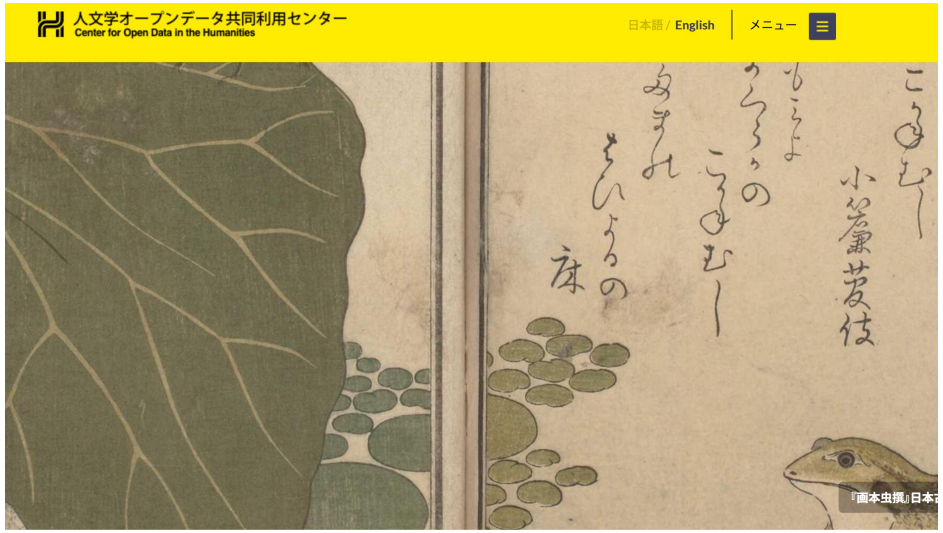
- 現在の資源
  - VU Amsterdam Metaphor Corpus (VUAMC) [Steen+10]
  - 日本語指標比喩データベース [加藤+20]（BCCWJベース）
  - 日本語レトリックコーパス (J-FIG) [小松原21]
- 利用方法
  - メタファー出現頻度の基準値 [Semino20]

NLPでの利用

- VUAMCの利用
  - メタファー検出 [Pramanick+18, Igamberdiev+18, 他多数]
  - クラウドソーシング利用手法改良のベンチマーク[Parde+17]
- マイナーな修辞表現のデータは独自に構築され、公開されず研究室内にとどまることも多い
  - キャッチコピーから抽出した対句のデータ [丹羽+20]

デジタル・ヒューマニティーズ

- 一般公開の事例
  - 人文学オープンデータ共同利用センター
  - 「日本古典籍データセット」「江戸料理レシピデータセット」など様々な人文学データを提供
- 修辞表現の実例も検索・閲覧できるとうれしい
  - これまで様々なレジスターで生み出されてきた名文・名言は日本語話者共通の財産



ROIS-DS人文学オープンデータ共同利用センター（Center for Open Data in the Humanities / CODH）は、情報学・統計学の最新技術を用いて人文学資料（史料）を分析する「データ駆動型人文学」や、人文学研究の成果に基づき構築したデータセットを超学際的に活用する「人文学ビッグデータ」など、オープンサイエンス時代の新しい人文学研究を展開します。[もっと詳しく..][CODHパンフレット..]

重要なお知らせ

2024-03-04
21th CODH Seminar - Digital History: Concepts and Practices

2024-02-29
【ROIS-DS】第4回成果報告会 データ駆動型研究の推進・支援活動

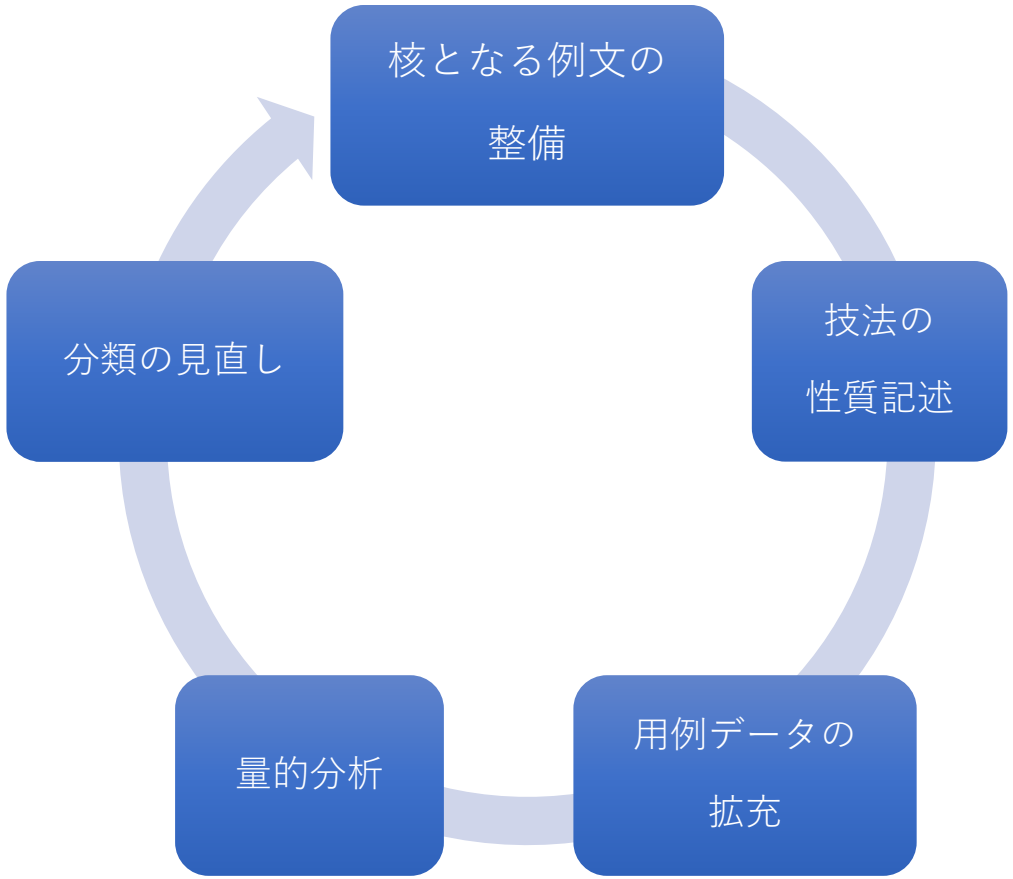
>> お知らせ一覧

(http://codh.rois.ac.jp/)

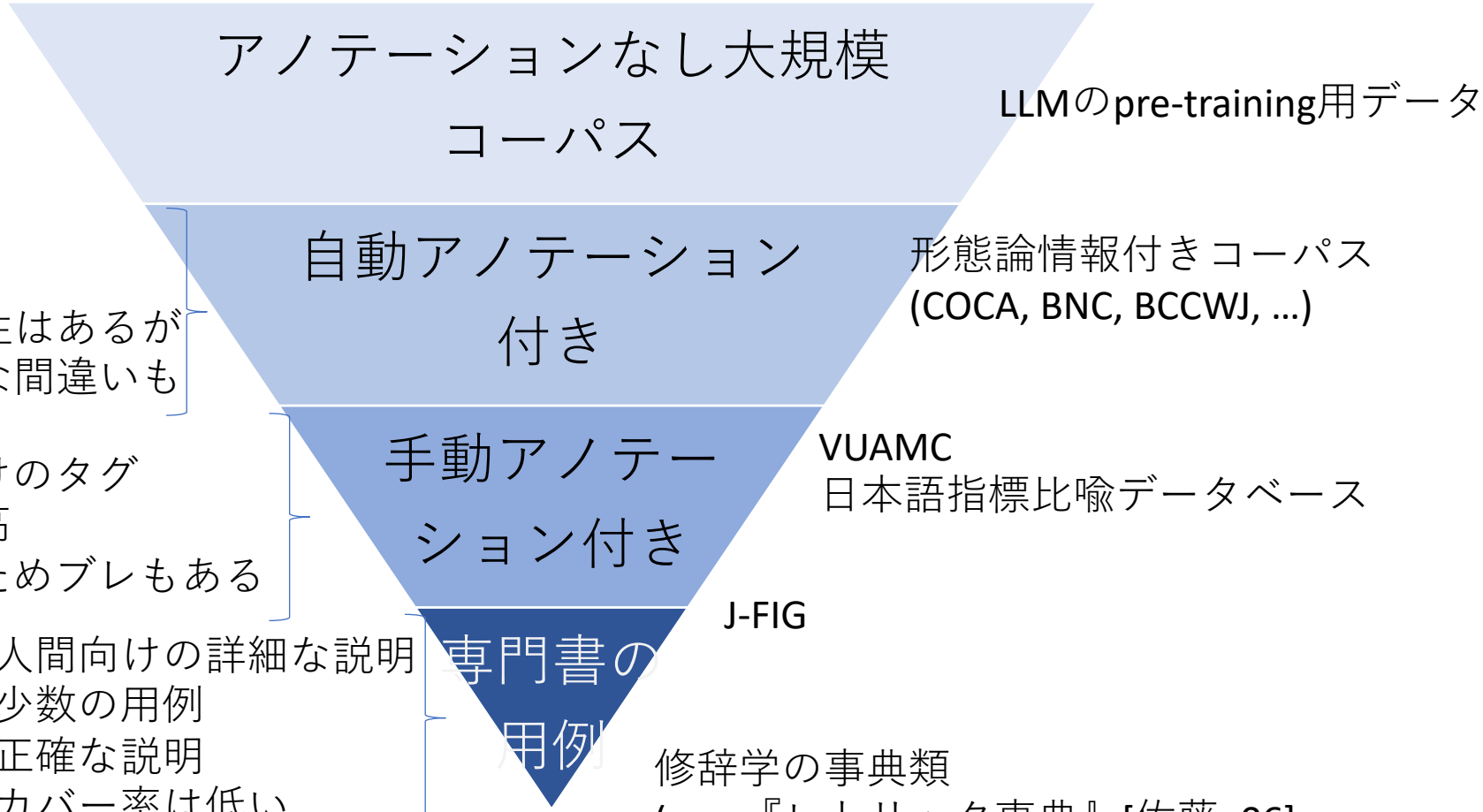
これからのレトリック
言語資源開発戦略を考える

言語学・修辞学とレトリック資源開発の研究サイクル

- 伝統的な言語学では少数の例文を対象
  - 典型的な事例による説明
- うまく説明できない事象も多い
  - 言語資源を構築するときに必ず当たる障壁
  - 現実は複雑
- 量的データを用いた研究の充実が必要
  - 科学としての言語学を洗練させる
  - 実証性の向上



言語資源拡大のプロセス

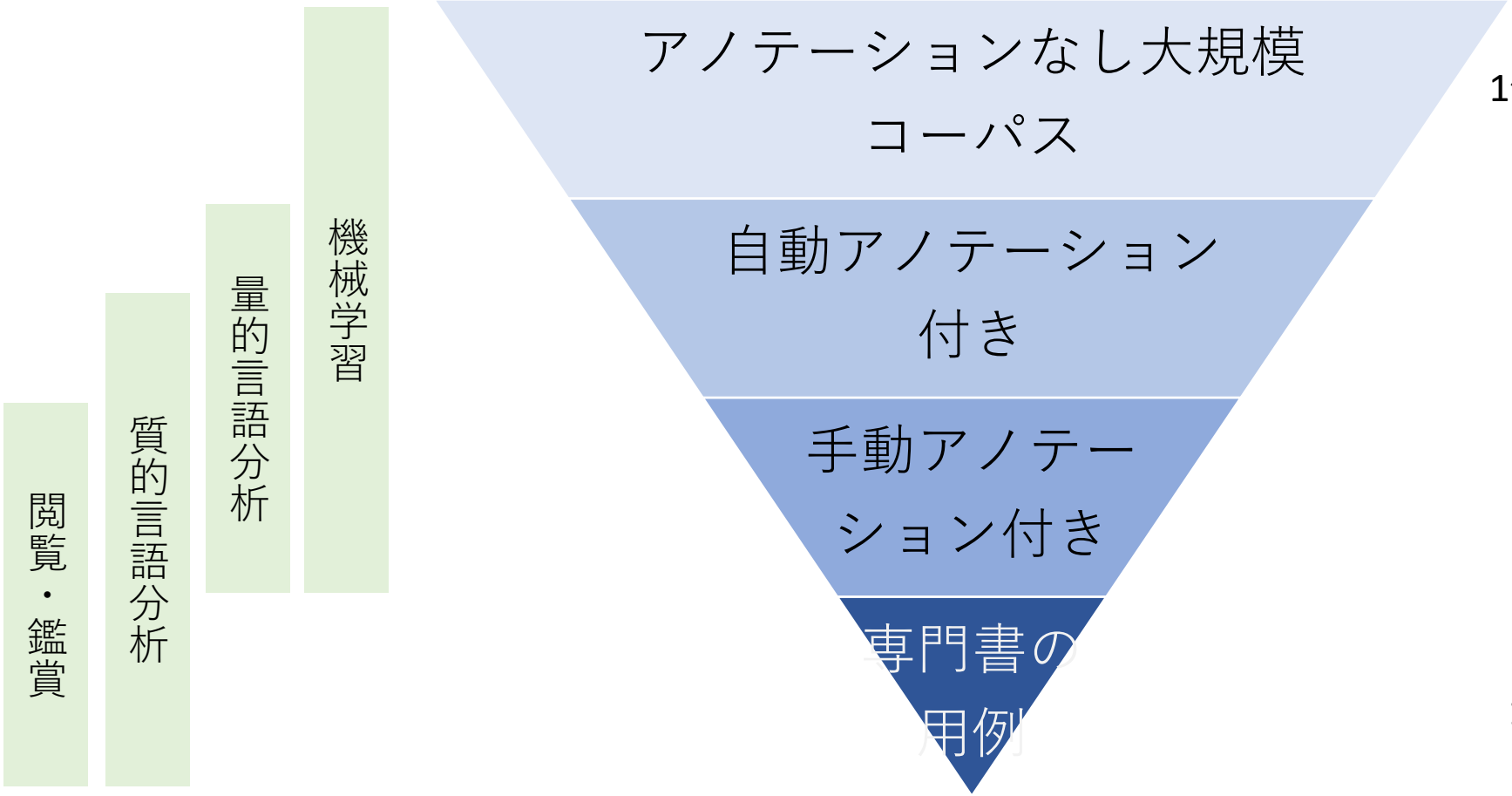


- 機械向けのタグ
- 規模が大きい
- アノテーションに一貫性はあるが
- 人がまず犯さないような間違いも

- 機械向けのタグ
- 信頼度高
- 人手のためブレもある

- 人間向けの詳細な説明
- 少数の用例
- 正確な説明
- カバー率は低い

言語資源における質・量のトレードオフ



分野による言語資源の傾向

- NLP
  - 性能・一貫性重視
  - 目的となるタスクにおいて、モデルの性能を上げることが目標
  - コーパスの量が（相対的に）重要
- 言語学
  - 理論的な正しさを重視
  - 人が検索、閲覧することを想定
  - クリティカルな用例や説明が重要
- 両者が相互乗り入れし、効率の良い言語資源の構築・使用をするにはどうすれば良いか？

要因1: NLPの動向

- fine-tuningの流行
  - LLMベースのAIはzero-/one-/few-shotで学習してしまう対象もある
  - これからの言語資源のサイズはどれだけ必要？
    - 学習に必要な例は以前より少なくて済む
    - 検証用データは依然として必要
  - 質的記述は重要になってくるか？
    - 辞書の定義文を用いた文埋め込みの作成 [塚越+23]
    - プロンプトエンジニアリングの隆盛
- 必要な資源の多様化
  - 既にあるメジャーな修辞表現のデータを拡大するのが有効？
  - マイナーな修辞表現のデータを新たに作成するのが有効？

要因2: 言語学での整備状況

- どれも人手でリッチな情報を付与しているが、カバーしている範囲に差
  - 修辞技法の種類
    - メタファー、メトニミー、シミリー（直喩）、アイロニー、オクシモロンなど
  - 言語の層
    - 意味論的側面、統語論的側面、語用論論的側面など
- 分類と解説
  - 分類：e.g. ある語がメタファーか否か
  - 解説：e.g. ある語がどのような概念領域に基づくメタファーで、どのような効果をもたらすか

VUAMC [Steen+10]

- BNC-babyの一部 (186,688語)に非メタファー(Non-MRW)、疑わしきは含める(WIDLII)、メタファー関係語(MRW)をアノテーション
- MIPVUと呼ばれる手続きで4レジスターを対象
- 概念メタファー理論 [Lakoff+80]に基づき、機能語のメタファーもアノテーション
- κ係数などの指標が提供されている

```
<text>
  <group>
    <text xmlns="http://www.tei-c.org/ns/1.0" xml:id="a1e-fragment01">
      <body>
        <div1 n="891002 edition -- Business and City Page 23" type="u">
          <head type="MAIN">
            <s n="1">
              <w lemma="late" type="AJS">Latest </w>
              <w lemma="corporate" type="AJ0">corporate </w>
              <w lemma="unbundler" type="NN1">unbundler </w>
              <w lemma="reveal" type="VVZ">
                <seg function="mrw" type="met" vici:morph="n">reveals</seg>
              </w>
              <w lemma="laid-back" type="AJ0">laid-back </w>
              <w lemma="approach" type="NN1">
                <seg function="mrw" type="met" vici:morph="n">approach</seg>
              </w>
              <c type="PUN">: </c>
              <w lemma="roland" type="NP0">Roland </w>
              <w lemma="franklin" type="NP0">Franklin</w>
```

日本語指標比喩データベース [加藤+20]

- BCCWJをベース
- 直喩の箇所に加え、概念マッピングなどの意味論情報、わかりやすさのアノテーションなど
  - VUAMC +αの情報量
  - BCCWJ由来の形態論情報などが利用可能

※同じセッションで発表があったばかりなのでスペック等の詳細は割愛させていただきます。

J-FIG [Komatsubara21]

- 様々な種類の修辞表現を収録
  - 164種類
- 総用例数は7,423
  - ロングテールの分布で、種類ごとの収録数にはばらつきがある
- 記述の幅が広い：「厚い記述」の方針
  - 修辞表現の種類、意味、構文、語用
- テキスト全体へのアノテーションではなく、焦点となる表現を中心にコンテクストと共に収録
  - 一般的な機械学習向けコーパスとは異なる
  - 人間が参照・閲覧することを前提
  - 一つの用例に様々な修辞表現に関する説明

「明子は彫塑のごとく佇めり」

Page Type	Example
Example ID	a0002
Author	芥川龍之介
Piece	「開化の殺人」
Reference	『芥川龍之介』
Pages in Reference	215

Text

「頭上の紫藤（しとう）は春日（しゅんじつ）の光りを揺りて垂れ、藤下の明子は凝然として彫塑のごとく佇めり。予はこの画のごとき数分の彼女を、今に至って忘るる能はず。」

Context	Focus	Standard	Context
	彫塑	明子	のごとく佇めり

Rhetoric

	Category
1	直喩・シミリ (simile)
2	擬物法・結晶法 (hypostatization)

（※数値は2024/3/8にhttps://www.kotorica.net/j-fig/へアクセスして確認したもの）

要因3: デジタル・ヒューマニティーズ

- デジタル・ヒューマニティーズとは
  - 「デジタル・ヒューマニティーズとは、人文学的問題を情報学的手法を用いて解くことにより新しい知識や視点を得ることや、人文学的問題を契機として新たな情報学の分野を切りひらくことなどを目指す、情報学と人文学の融合分野である。またデジタル・アーカイブはデジタル・ヒューマニティーズの成果公開の有力な方法の一つである。」
    (http://agora.ex.nii.ac.jp/~kitamoto/research/dh/)
- 非常に広い定義だが、言語学のコーパス利用もこの一部と捉えられる
- 修辞表現研究における応用可能性
  - 用例をリンクで結び、用例間の関係ネットワークを記述できる

暗示引用、パロディ

- いわゆる「元ネタ」があって成立する修辞表現
- 夏目漱石『草枕』
  - 山路を登りながら、こう考えた。
    智に働けば角が立つ。情に棹させば流される。意地を通せば窮屈だ。とかくに人の世は住みにくい。
    住みにくさが高じると、安い所へ引き越したくなる。どこへ越しても住みにくいと悟った時、詩が生れて、画が出来る。
- 井上ひさし『おれたちと大砲』
  - （…）その、羊腸の如くくねった、おれはこう考えた。
    薩長の反逆を思えば腹が立つ。君家の窮状を思えば涙が流れる。腹立ちと涙を押えて暮らすのは窮屈だ。とにかくに人の世はお先まっくらだ。お先くらいのが高じると、明るい所へひっ越したくなる。どこへ越してもくらいと悟った時、おれたちのように、戦おうとするものが生まれる……
- デジタルだとネットワークを表現しやすいし、扱いやすい

LLMは言語資源構築にどう影響するか

- これからの機械学習にどの程度のデータ量が必要か？
- LLMは言語学の専門書を理解できるか
  - 分類モデルへの応用
  - 自然言語理解の進展
- LLMは新しい用例の生成や検索に利用できるかもしれない
  - 不安や使いづらい面も
    - 生成される用例のバイアス
    - メタ情報や使用文脈
- **人間が使う言語資源と機械が使う言語資源の接近**

これからのレトリック言語資源構築戦略

- 生テキストの共通化と記述の多層化
  - テキスト収集、著作権処理などのコストを圧縮
  - 共通のテキストにアノテーションすることで多様な情報を利用可能
    - あらかじめ施されている形態論情報や、他の研究で施されたアノテーションの利用
  - BNC, COCA, BCCWJなど、現在基盤となっている言語資源が確立してからできた**SNS**など、、新しいレジスターのデータも追加していく必要も
  - 用例間の関係の記述
    - 特に（明示的にせよ暗示的にせよ）引用
- 用例数は（今まで想定されていた最大よりも）少なくて済む？
  - 少ない用例で機械学習できるようになっても性能検証できる程度の量は最低限必要
- 用例あたりの記述の質・量の向上
  - モデルが人間向けの記述を「理解」できるようになるのであれば、人間向けの記述も役立つ
- カバーする修辞表現の種類を広げる
  - 生成される文体の操作

おわりに

- 思索的で不確定要素も多く、明確な根拠に基づく発表ではありませんが、今後の言語資源のあり方や構築戦略に関する議論につながれば幸いです。

ご清聴ありがとうございました

謝辞

本研究は国立国語研究所基幹型プロジェクト「実証的な理論・対照言語学の推進」・サブプロジェクト「アノテーションデータを用いた実証的計算心理言語学」によるものです。
また、本研究は JSPS 科研費 23K12164 の助成を受けたものです。