

自然言語処理のタスクを 日本語史研究に落とし込む

北崎勇帆（きたざき ゆうほ）

高知大学人文社会科学部

yuho@kochi-u.ac.jp

自己紹介と発表の構成

- 日本語学 > 日本語史 > 日本語文法史

1. 日本語史研究の紹介

- ちょっとした事例紹介
- 研究の主な目的
- 日本語史研究者が知りたいこと

2. NLPのタスクとの照らし合わせと概観

自己（の事例）紹介

1. 意志の（ヨ）ウが理由節に生起できなくなる（北崎2021）
 - みんなおめへに。**あげやう**から。よんでみな（花街鑑 [1822刊]）
 - みんなお前に {*あげようから／あげるから／あげよう。だから}、読んでみな。
2. 願望のタシ（>タイ）は文末では使われにくかった（北崎2022）
 - 西国へモ**仕タク**候シカドモ、八条院へ兼参ノ身ニテ候シ時ニ（延慶本平家物語 [13C]）
 - 宇治川を渡いた事と、源三位入道の討ち死にを召された所をも**聞きたい**（天草版平家物語2-6 [1592刊]）

何を目的とするか？

- 共時的な側面の記述
- 通時的な側面の記述
 - いつ {成立／衰退／交替} したか、何に、どのような側面（音韻・形態・意味・機能…）で起こったか。
 - どのような地域や文体、話者・集団に起こったか。

何を目的とするか？

- その一般化や要因の解釈
 - どのように（どのようなプロセスで）変化したか？
 - なぜその変化が起こったのか？
 - cf. Harris and Campbell 1995, 渋谷2008, 堀田2015, 小柳2018
- さきほどの事例でいうと、
 - 「あげようから」の衰退：終止形のテンス的未来への進出
 - 願望の事例：「ご飯を食べたい！」の領域を終助詞と分担している

何が分かる と 嬉しいか？

- 目的の裏返しとして、
 - 言語変化が起こっていることと、その記述
 - 「これって現代語では言わないけど、江戸時代には普通にあるね」
 - 共時的なスタイル・レジスター差
 - 「書き言葉寄りの資料には出るけど、話し言葉の資料には出ないね」
- 取っ掛かりは経験・知識に基づく直感頼り
 - 資料を読みながら、他のデータを見ながら、etc.
 - 現代語研究・内省と対照しながら
 - 必然的に、仮説検証によるミクロな記述の積み重ねから一般化していく方向性に
- 「どのように」「なぜ」も分かるようになる？
 - (Kuteva et al. 2019, Narrog and Heine 2021)

タスクを落とし込む

- NLPのタスクを日本語史研究の文脈で考えてみる
 - [NLP-Progress](#) と [その解説](#) を参考に
- 系列ラベリング
 - 品詞タグ付け、構造解析・構文解析
 - 固有表現抽出
 - データが有限のようで、そうでもなくなってきた？
 - NDL-OCR: https://lab.ndl.go.jp/data_set/ocr/r3_text/
 - NDL古典籍OCR: https://lab.ndl.go.jp/data_set/r4ocr/r4_koten/

タスクを落とし込む

- 語義曖昧性解消
 - 現状のCHJには意味・用法のアノテーションはついていない (→[P4-10](#), [P3-12](#), 小木曾2019)
 - ただし、意味・機能を一意に決定できるという前提はやや怪しい
 - e.g. 格助詞・接続助詞の「に」「を」「が」 (近藤2000)
 - 未知語義推定・記述、変化の検出がいけるともっと面白そう (→[T2](#), [P4-8](#), 相田ほか2020)
 - 意味をどう切り分けるかは必ずしも自明ではない

タスクを落とし込む

- 文書分類
 - 資料ごとの文体価値（和文 VS 漢文訓読文）の定量的評価（築島1963, 大川2020）
 - ジャンル・レジスター差はここ最近のトレンドでもある（cf. [日本語文法22\(2\)](#)）
- 文法誤り訂正（GEC）
 - 次で詳しく
- 生成系は落とし込みにくいかも

文法誤り訂正による変化の検出

- 時代Aを基準とした時代Bとの差分を「誤り」と捉えてみる
 - 形式の有無のレベルで
 - 発生：是**ば**しいだし参らすな（高野本平家物語・巻6 [13C]）
 - 衰退：我妹子をいざみの山を高**み**かも大和の見えぬ国遠**み**かも（万葉集1-44 [8C]）
 - 交替：[奈良] まくほし→[平安] まほし→[鎌倉] たし（→たい）
 - 意味・用法のレベルで
 - 我が恋は千引きの石を七**ばかり** [=くらい] 首に掛けむも神のまにまに（万葉集4-743 [8C]）
 - 月影**ばかり** [=だけ] ぞ、八重葎にもさはらずさし入りたる。（源氏物語・桐壺 [1010 頃]）
 - 組み合わせのレベルで
 - [～平安] *死ぬ+ぬ（<往ぬ）→[鎌倉] 死にぬ

文法誤り訂正による変化の検出

- 「言えるけど不自然」
 - ソコヲハカラウ尉ノ官ノ者ガハラタツテ、「**タレナレバ**夜_ルトヲルゾ」ト云タゾ（玉塵抄 [1563]）
 - ??あなたは誰 {なので/だから} ここを通るんですか？
 - どういう理由があって/どういう理由で/どういう権限があって
 - 流暢性のレベル？ → [Q1-9](#)

文法誤り訂正による変化の検出

以下の会話文は、誤りを含んでいます。正しい日本語の会話文に修正したものを出力した後、修正した箇所とその理由を箇条書きにして示してください。（ChatGPT）

- 主「米八じゃあねえか。どうして来た。そして隠れて居る此所が知れる**というも**ふしぎなこと。まあまあこちらへ、夢じゃあねえか」と、おきかえりてすわる。
- よね「**わちきゃあ**もう、知れ**めえ**かと思って胸がドキドキして、そしてもう急いで**歩行た**もんだから、あ、苦しい」とむねをたたき「喉がひつつくようだ」といひながらそばへすわり、
- 「**おまはん**は煩ってい**さっしゃる**のか**え**」とかおをつくづく見て「**寔に**やせたねえ。マア色のわりいこと**わ**。真青だよ。何時分からわるいのだ**え**」
- 主「ナ二十五六日**跡**からよ。**大造な**ことでもねえが、どうも気が閉でならねえ。それはいいが手めえまあ、どうして知って来たのだ。聞てえこともたんとある」

（春色梅児誉美 [1833刊] ）



- 主：「米八じゃあねえか。どうして来た。そして隠れているところが知られる**という、も**ふしぎなこと。まあまあこちらへ、夢じゃないか」
- よね：「**私**、もう知られたかと思って胸がドキドキして、急いで**歩いて**きたんです。あ、苦しい。」と**言いながら**、胸をたたき「喉が渴くようだ」と近くに座る。
- 「**あなた**は心配なさっているの？」と、彼をじっと見つめて「**本当に**痩せましたね。まあ、色のわりには悪くないですけど、真青ですよ。何日から悪くなったの？」
- 主：「何十五六日**前**だろう。**大した**ことではないんだけど、どうも気になって仕方がないんだ。それはともかく手元に来てくれたことに感謝しているよ。聞きたいことがいろいろあるんだけど。」

文法誤り訂正による変化の検出

- 修正した箇所とその理由：
 - 「此所」 → 「ここ」：「此所」は古風な言い回しで、現代の日本語では「ここ」を使います。
 - 「もふしぎなこと」 → 「なんとも不思議なことだ」：「もふしぎなこと」は聞き慣れない表現で、意味が分かりにくいため、「なんとも不思議なことだ」と言い換えることで、より明確に表現します。
 - 「～というのも」の準体句への違和感までは抱けている！
 - 「夢じゃあねえか」 → 「夢じゃないか」：「夢じゃあねえか」は方言で、標準語では「夢じゃないか」と言います。
 - 「わちきゃあ」 → 「私は」：「わちきゃあ」は方言で、標準語では「私は」と言います。
 - 「知れめえかと思って」 → 「知られたかと思って」：「知れめえか」という表現は間違いで、正しくは「知られたか」と言います。
 - 「歩行たもんだから」 → 「歩いてきたんです」：「歩行たもんだから」は不自然な表現で、正しくは「歩いてきたんです」と言います。

文法誤り訂正による変化の検出

- ただし…
 - 正解一誤りの対応付いた教師データがない
 - 擬似誤り生成で対処可能？ (Yasunaga, Leskocev and Liang 2021)
 - 定量的な評価ができない
 - 本当に（当時の規範として）「誤り」と言えるかがそもそも分からないという根本的な問題
 - 「変化と言えそう」というところにゴールを下げて、「言わないよね」が分かる人を連れてくる？ (小田2015)

文法誤り訂正による変化の検出

- 翻訳も同様のアプローチに還元可能
 - 江戸前期→大正の俗語訳の事例（北崎2020）
 - a. ▲瓜盗「…まをし，ものを仰らねば何とも迷惑でござる。重ねては最早参ります**まいほどに**，平にお許させられて，返させられてくだされませや。
（続狂言記・巻2 [1700 刊]）
 - b. 瓜盗『…もし，さう黙つてみられては，怖しう御座います。之からは決して参りませ**んから**，どうか御免下さい。（新訳狂言記 [1919 刊] 瓜盗人）
- 共時的なレジスター差にも適用可能かも？
 - 「平安時代の漢文訓読文」 VS 「平安時代の和文」

おわりに

- 従来のな日本語史研究のアプローチ
 - 人間が発見してデータで実証
 - 言語処理はあくまでも補助的な位置付け
- 問題発見の過程にNLPが関与してくる時代が来るはず
 - 宣伝：
日本語学会2023年度春季大会シンポジウム（2023/5/21）
「情報技術と大規模テキスト資源がひらく日本語史研究」
<https://www.jpling.gr.jp/taikai/2023a/>

参考文献

- Harris, A. C., & Campbell, L. (1995). *Historical Syntax in Cross-Linguistic Perspective*. Cambridge University Press.
- Narrog, H., & Heine, B. (2021). *Grammaticalization (Oxford Textbooks in Linguistics)*. Oxford University Press.
- Kiyono, S., Suzuki, J., Mita, M., Mizumoto, T., & Inui, K. (2019). An empirical study of incorporating pseudo data into grammatical error correction. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019* (pp. 1236-1242). Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/D19-1119>
- Kuteva, T., Heine, B., Hong, B., Long, H., Narrog, H., & Rhee, S. (2019) *World Lexicon of Grammaticalization. 2nd Edition*. Cambridge University Press.
- Yasunaga, M., Leskovec, J., & Liang, P. (2021). LM-Critic: Language Models for Unsupervised Grammatical Error Correction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 7752-7763). <http://dx.doi.org/10.18653/v1/2021.emnlp-main.611>
- 相田太一・小町守・小木曾智信・高村大也・坂田綾香・小山慎介・持橋大地（2020）「単語分散表現の結合学習による単語の意味の通時的変化の分析」『言語処理学会 第26回年次大会 発表論文集』 pp.485-488. https://www.anlp.jp/proceedings/annual_meeting/2020/pdf_dir/E2-3.pdf
- 大川孔明（2020）「文連接法から見た平安鎌倉時代の文学作品の文体類型」『日本語の研究』 16(2), pp.133-151. https://doi.org/10.20666/nihongonokenkyu.16.2_133
- 小木曾智信（2019）「『日本語歴史コーパス』への追加情報の付与と共有 —中古和文の「る」「らる」を例に—」日本語学会2019年度春季大会. https://www.jpling.gr.jp/taikai/happyo/yosi/2019/a/yosi_2019a_h1.pdf
- 小田勝（2015）『実例詳解古典文法総覧』和泉書院.
- 北崎勇帆（2020）「近代に口語訳された狂言記」『国語語彙史の研究39』和泉書院, pp.117-136. (PDF)
- 北崎勇帆（2021）「中世・近世における従属節末の意志形式の生起」『日本語の研究』 17(2), pp.19-36. https://doi.org/10.20666/nihongonokenkyu.17.2_19
- 北崎勇帆（2022）「希望表現の史的変遷—願望を中心に—」青木博史・岡崎友子・小木曾智信（編）『コーパスによる日本語史研究 中古・中世編』ひつじ書房, pp.109-131. (PDF)
- 小柳智一（2018）『文法変化の研究』くろしお出版.
- 近藤泰弘（2000）『日本語記述文法の理論』ひつじ書房.
- 渋谷勝己（2008）「新たなことばが生まれる場」金水敏・乾善彦・渋谷勝己『シリーズ日本語史4日本語史のインタフェース』岩波書店, pp.101-138.
- 築島裕（1963）『平安時代の漢文訓読語につきての研究』東京大学出版会.
- 堀田隆一（2015）「言語変化研究の多様な視点—5W1Hによる整理—」『人文研紀要（中央大学）』 81, pp.293-319. <http://id.nii.ac.jp/1648/00007282/>