言語資源小委(R6.7.11)
資 料 5

日本語の言語資源
特に言語コーパスについて

前川喜久雄（国立国語研究所）

今日のプレゼンテーションの内容

1. 言語資源／言語コーパスとは
2. 日本語コーパスの整備
3. 現状分析と今後の課題
4. まとめ

おことわり

以下の内容は私の個人的見解です。国立国語研究所の公的な見解を述べるものではありません。また私は言語資源開発の現場を離れて 7 年ほどたちますので、一部の内容は最新の状況をとらえきれていないかもしれません。もしお気づきの問題があればご指摘ください。

言語資源とは何か

および自然言語自体

事典『ウィキペディア（Wikipedia）』

言語資源（げんごしげん、英: Language resource）とは、自然言語を研究するさいに用いられる資源のこと。 辞書やコーパス、シソーラス、インフォーマントなどがこれにあたる。 電子化された言語資源は自然言語処理技術の研究に不可欠であるが、作成に非常に手間がかかるため、いまだにその数は少なく、一般にとても高価である。

一部の特権的な組織しか大量に備蓄できない
従って公的組織の関与が期待される

近年WWWが普及したこともあり、これらの資源をインターネット上から自動的に獲得しようとする試みも数多くなされてはいるが、一般的なネットワーク上の文章にはノイズが多すぎて価値ある情報を収集するのは難しいとされる。

LLM開発におけるBCCWJのフィルターとしての利用

また、言語資源には著作権の問題が重くのしかかっている。それはたとえ資源を作っていても、それを公開するのは権利上の許可を得なければならないからである。 ウィキペディアはこの問題に対する解決策の一つとなるべく運営されている。

今日は触れないがきわめて重要

言語資源の具体例

- **言語コーパス**

  実際の言語活動を体系的に記録し電子的に検索可能として公開した言語資料

- **解析用電子化辞書**

  コーパスに品詞などの形態論情報を自動付与するための解析用辞書

- **オントロジー**　概念分類体系

- **アノテーションスキーマ**

  テキスト、音声、映像などに検索用情報を付与するための体系と作業原則

- **標準化規格**　データ相互利用のための標準化規格

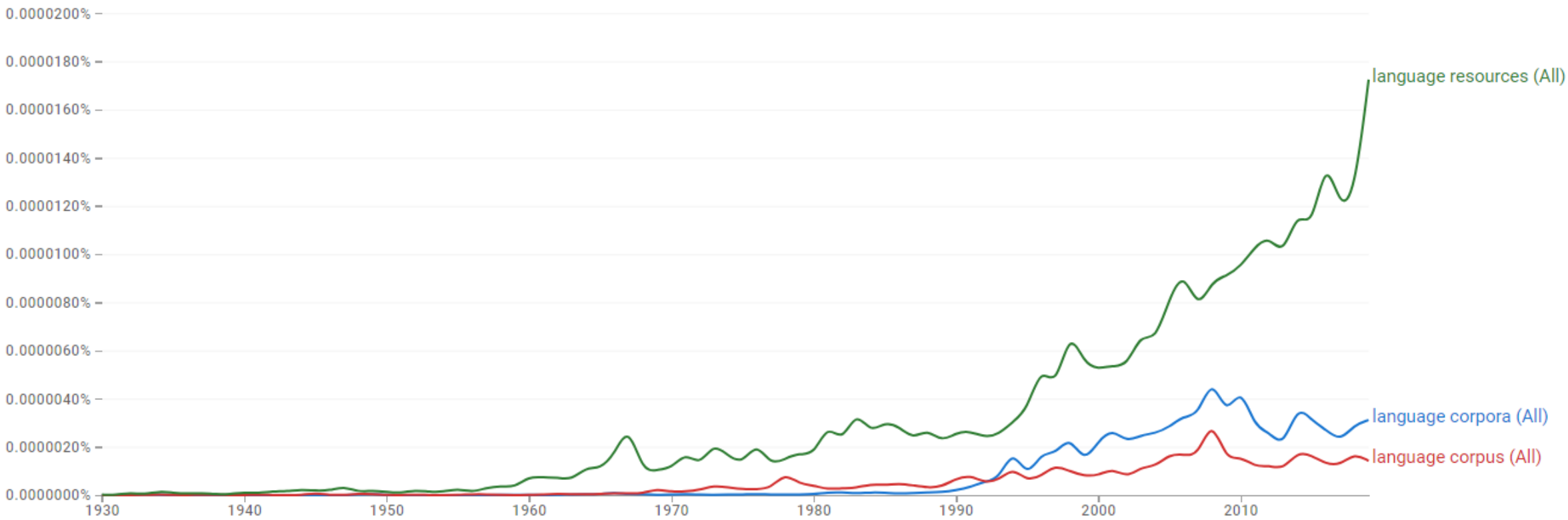
- **解析用ソフトウェア**

  音声認識、形態素解析、構文解析、分散表現、attention分析、etc.

Google Books Ngram Viewer

language corpora,language corpus,language resources

1930 - 2019 | English (2019) | Case-Insensitive | Smoothing of 0



(click on line/label for focus, right click to expand/contract wildcards)

言語コーパスの要件（前川2013）

- **真正性**　実際に生じた言語活動の記録であること
- **均衡性**　対象言語の様々なレジスターをカバーしていること
- **代表性**　母集団の縮図であること（可能でないことも多い）
- **規模**　利用目的に応じたデータ量を提供できること
- **電子化**　電子的な検索が可能であること
- **付加情報**　電子的な検索に必要な情報が付加されていること
- **公開**　誰もが利用できること（必ずしも無償を意味しない）

言語コーパスの要件（最低限）

- **真正性**　実際に生じた言語活動の記録であること
- **均衡性**　対象言語の多くのレジスターをカバーしていること
- **代表性**　母集団の縮図であること（可能でないことも多い）
- **規模**　利用目的に応じたデータ量を提供できること
- **電子化**　電子的な検索が可能であること
- **付加情報**　電子的な検索に必要な情報が付加されていること
- **公開**　誰もが利用できること（必ずしも無償を意味しない）

コーパスの種類

- 均衡コーパス　　レジスターの広がりを重視
- サンプル／全文コーパス　　サンプリングの有無
- モニターコーパス　　継続的にアップデート
- 共時／通時コーパス　　ひとつの時代だけか歴史的か
- 相同コーパス　　同じ設計で複数言語のコーパスを
- 対訳コーパス　　原文と翻訳の対
- 学習者コーパス　　非母語話者による言語活動

コーパスの類似概念

- アーカイブズ

  限定された対象のすべてを保管。メタデータ以外のアノテーションは施されないことが多い（先述のコーパスの最低限の要件にほぼ該当）

  国会会議録、地方議会議事録
  青空文庫（17557作品）
  国立国会図書館デジタルコレクション（図書・雑誌 約25万件）

- 言語データベース

  コーパスの要件のいくつか（特に真正性）を前提としないデータ、あるいはコーパスなどを加工して得られた二次データ

  朗読音声データベース
  漢字データベース
  単語頻度表
  単語共起頻度データベース
  etc.

コーパスの必要性：言語研究

- 言語に関する我々の内省は不十分
  「山がそびえる」「撤回しないべきだ」「気づかなそうな」
- 言語とそれが用いられる状況との間に相互作用がある
  待遇表現（敬語）、レジスター差（言語の使用目的による差）
  状況を周辺化したとき言語が機能しうるかは経験的に検討すべき問題
- 言語には創造性があり、また時間とともに変化する

  新語・新表現・外来語・集団語、伝統的表現の意味変化
- Phraseologyの必要性

コーパスの必要性：情報科学

- 1980年代まで主流であった記号ベース（規則ベース）のアプローチが1990年代以降、統計ベース（学習ベース）のアプローチにとってかわられるようになった

  ⇒　音声認識や音声合成の実用化

- 2010年代になって深層学習が登場し、この傾向は一層顕著になった

  ⇒　機械翻訳、大規模言語モデル

コーパスの必要性：潜在的文化財

- 言語データの大部分は時間とともに失われる
- 特に話し言葉に顕著だが実は書き言葉も同様
- 言語コーパスは時間の経過とともに文化財的な価値を増す

コーパスの必要性：科学的言語政策の礎

これから日本語は社会の変化によって大きな変化期をむかえることが予想され、それに応じた教育上・文化上の施策が必要になる。
その施策を科学的な根拠に立脚させるためにデータが必要

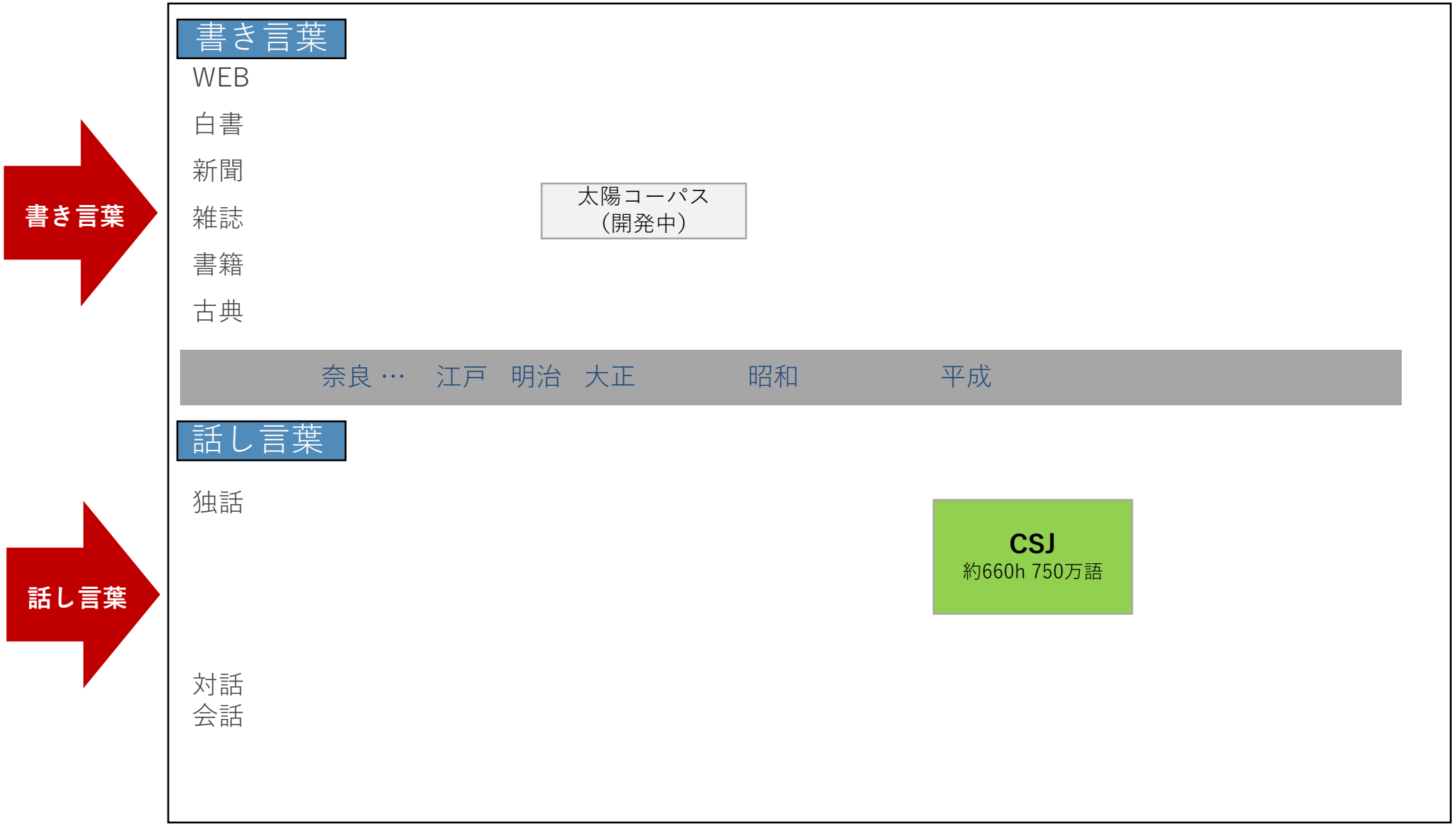
- **サイバー化**　　話し言葉・書き言葉　＋　打ち言葉
- **高齢化**　　コミュニケーション補助
- **多文化・多言語化**　　社会の3-4%が日本語を母語としていない
- **AIによる知的労働の代替**　　「外国語ができる」の意味が変化

のちほど別のスライドでもう少し詳しく

日本語コーパスの整備

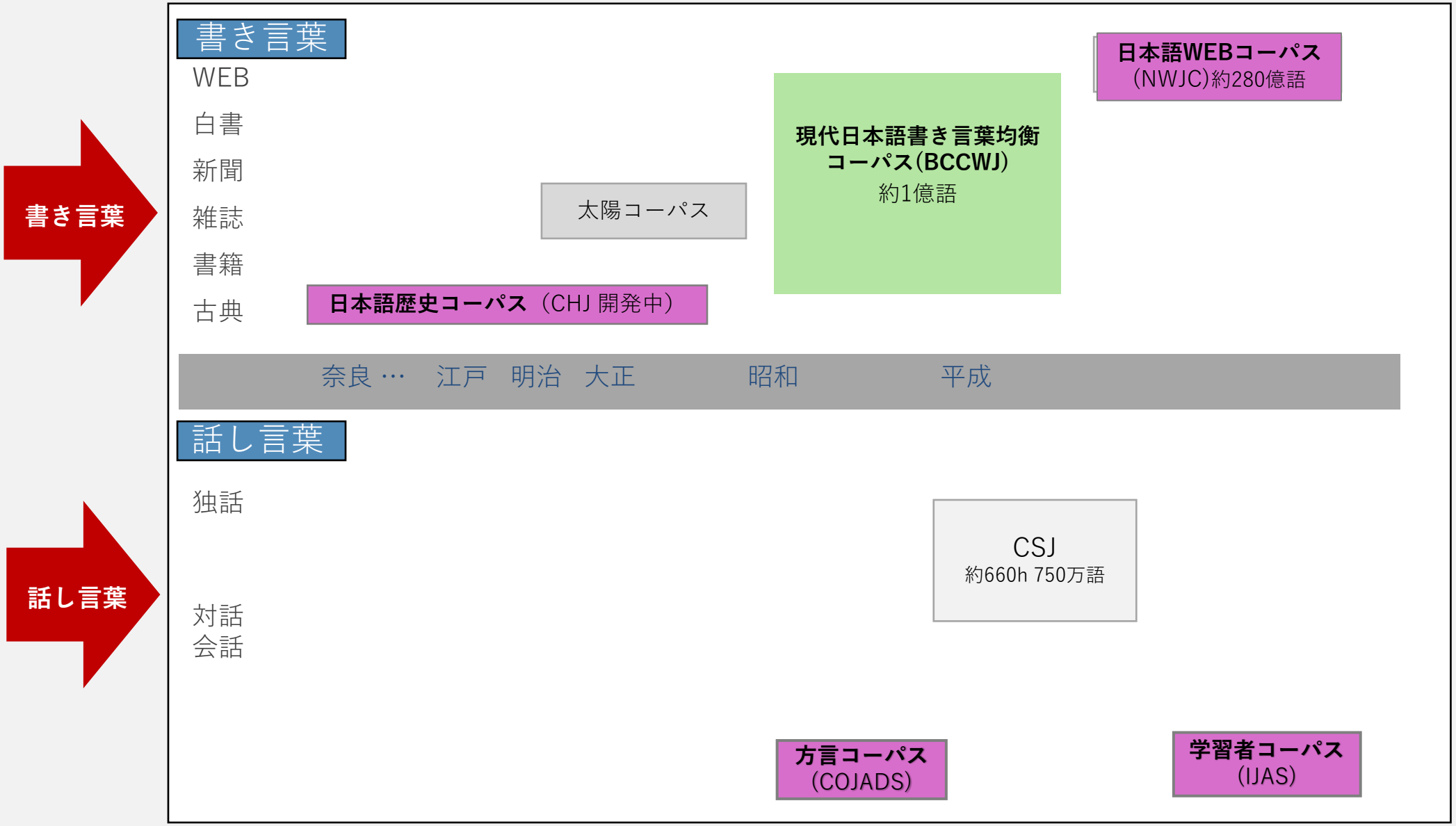
国語研での開発

『日本語話し言葉コーパス』CSJの構築(1999-2003)

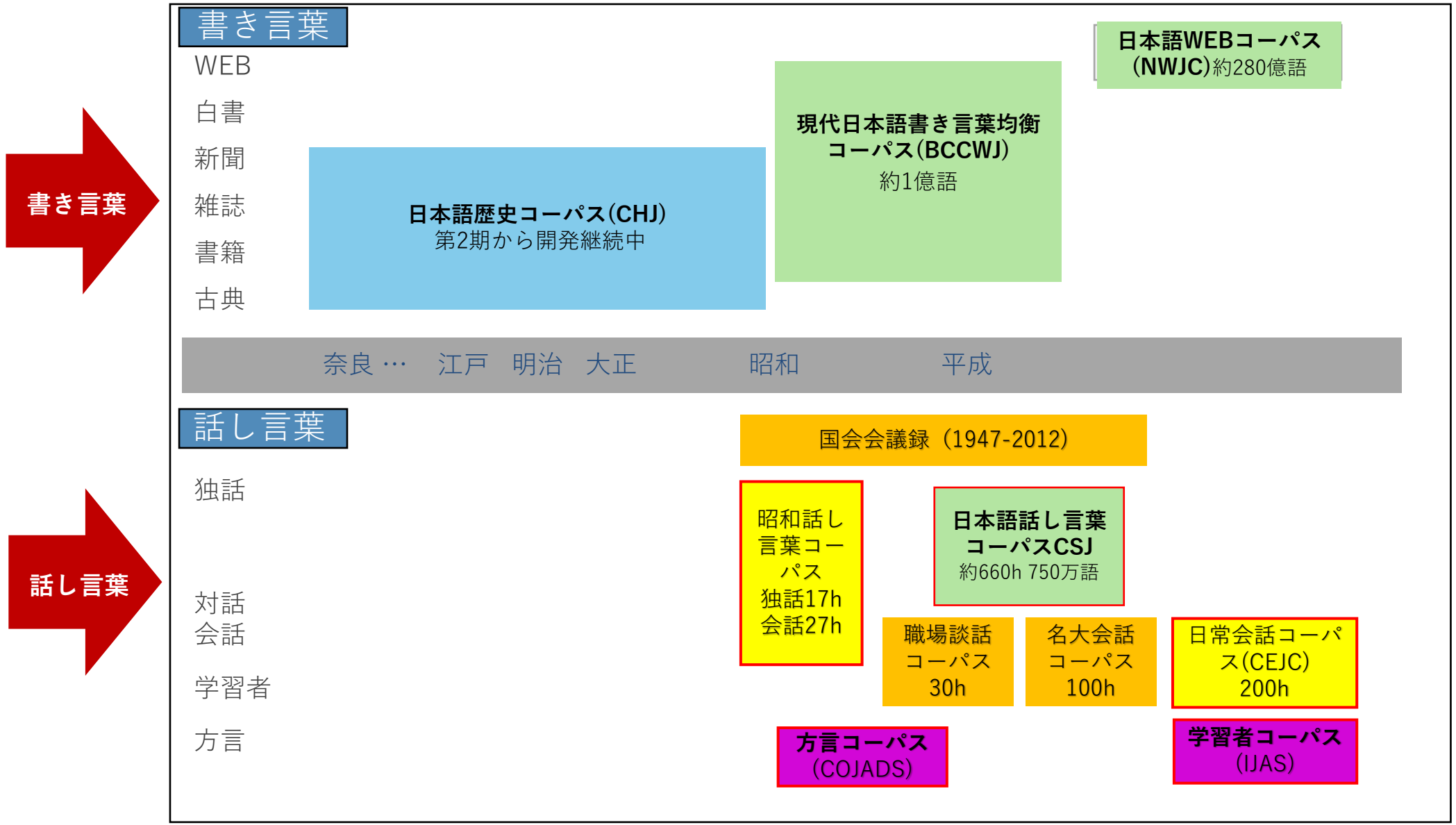


BCCWJの構築(2006-2010)

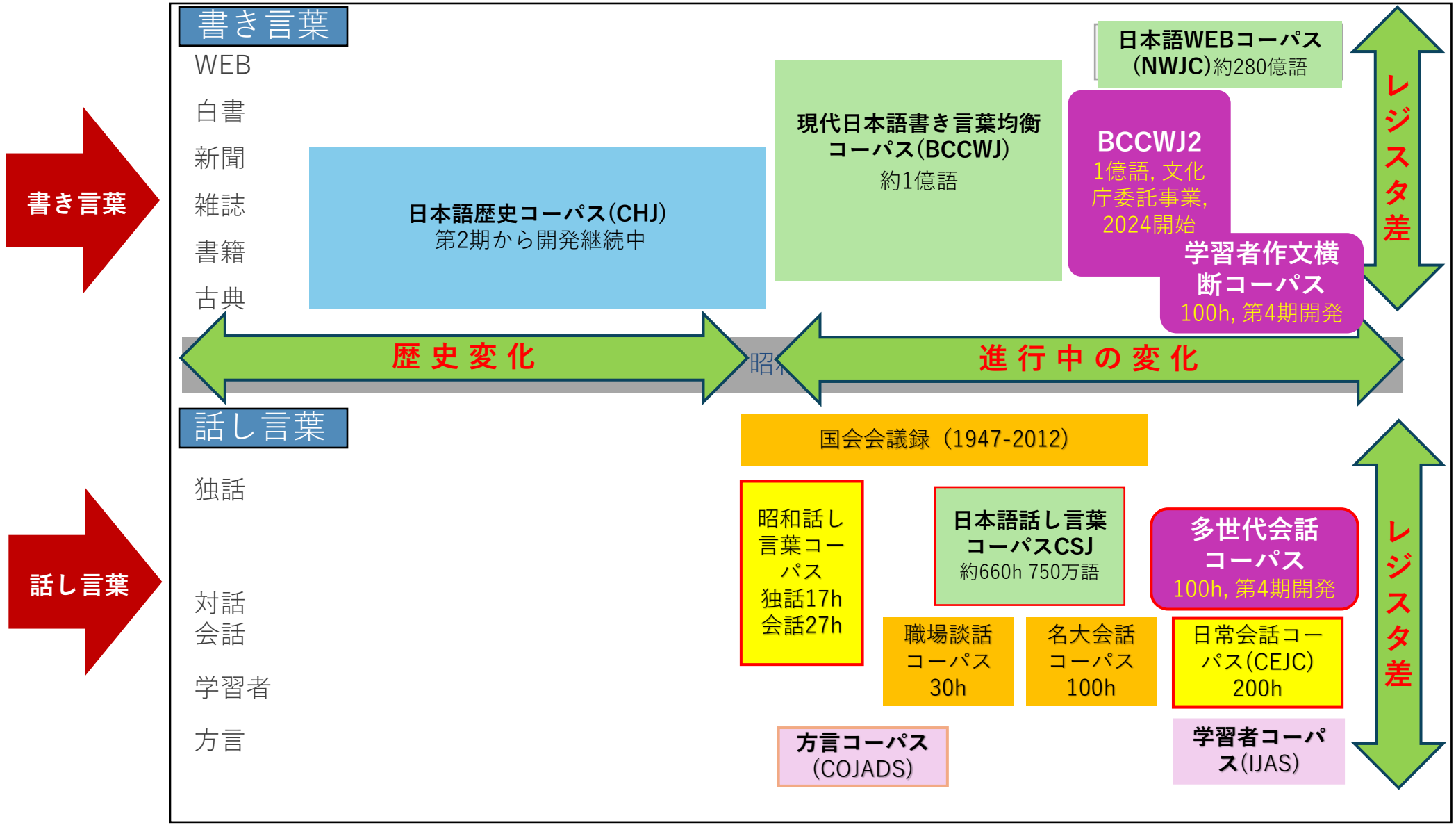
第2期中期計画期間(2010-2015)末



第3期中期計画期間(2016-2021)末



現在進行中


書き言葉
WEB
白書
新聞
雑誌
書籍
古典
日本語歴史コーパス(CHJ)
第2期から開発継続中
現代日本語書き言葉均衡コーパス(BCCWJ)
約1億語
日本語WEBコーパス(NWJC)約280億語
BCCWJ2
1億語, 文化庁委託事業, 2024開始
学習者作文横断コーパス
100h, 第4期開発
レジスタ差
書き言葉
歴史変化
昭和
進行中の変化
話し言葉
独話
対話
会話
学習者
方言
国会会議録（1947-2012）
昭和話し言葉コーパス
独話17h
会話27h
日本語話し言葉コーパスCSJ
約660h 750万語
多世代会話コーパス
100h, 第4期開発
職場談話コーパス
30h
名大会話コーパス
100h
日常会話コーパス(CEJC)
200h
方言コーパス(COJADS)
学習者コーパス(IJAS)
レジスタ差
話し言葉

国語研外のコーパス/データベース（ごく一部）

- 日本音響学会新聞記事読み上げデータベース(JNAS)
- 京都大学テキストコーパス
- 京都大学格フレーム
- 日本語係り受けコーパス（京都大）
- Web日本語Nグラム(Google)
- TWC (Tsukuba Web Corpus)
- Universal dependencies Japanese BCCWJ
- NICT JLE (Japanese Learner English) Corpus
- International Corpus of Learner English
- CHILDES

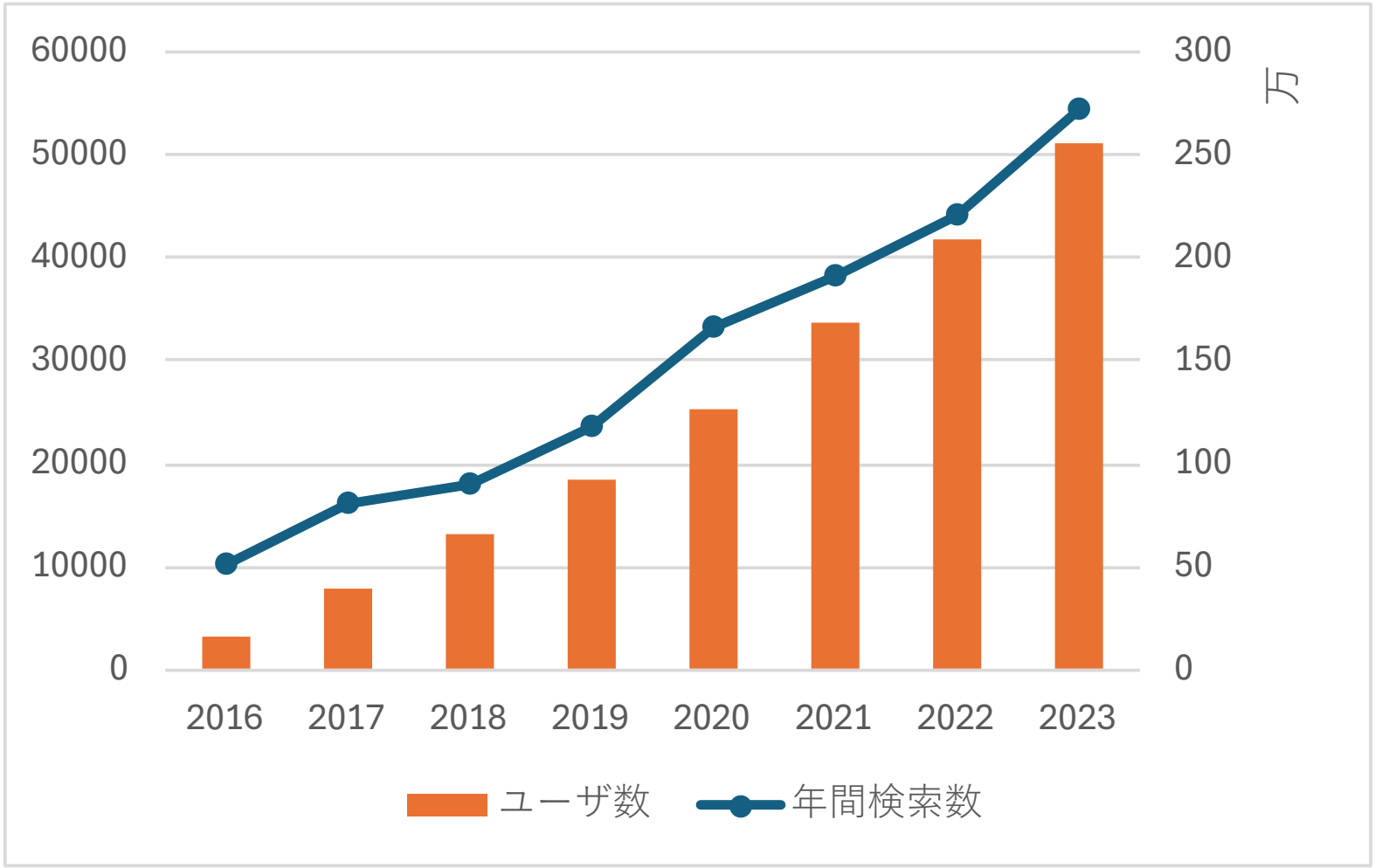
言語資源配布機関（日本国内の公的組織）

- **GSK（言語資源協会）**
  言語資源全般（無償／有償 計39種）
  https://www.gsk.or.jp/

- **ALAGIN（高度情報融合フォーラム）**
  NICT（情報通信研究機構）開発のデータ（会員限定16種）
  https://alagin.jp/

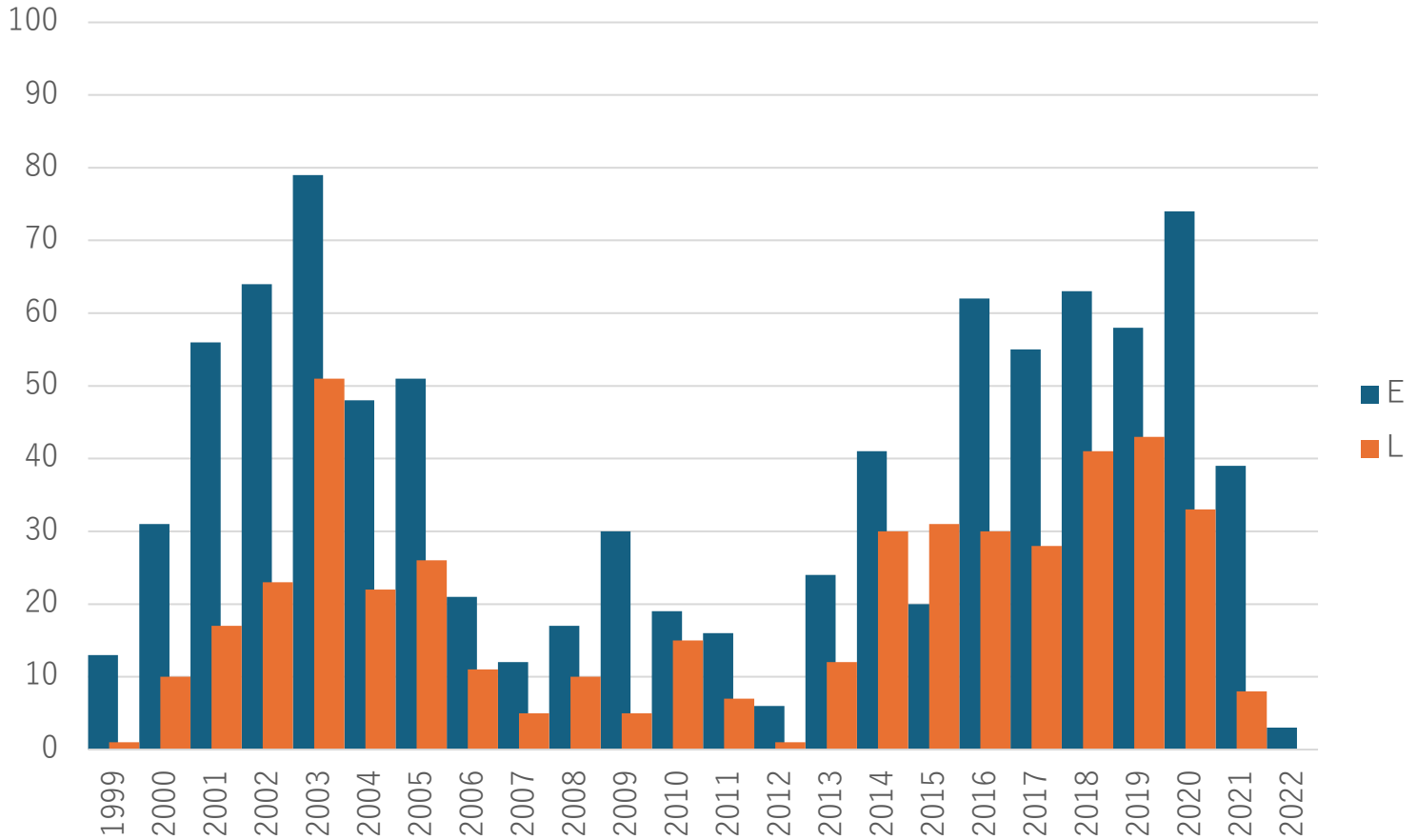
- **NII-SRC（音声資源コンソーシアム）**
  音声関係全般（無償36種、有償4種）
  https://research.nii.ac.jp/src/

言語資源群の活用状況（国語研の場合）



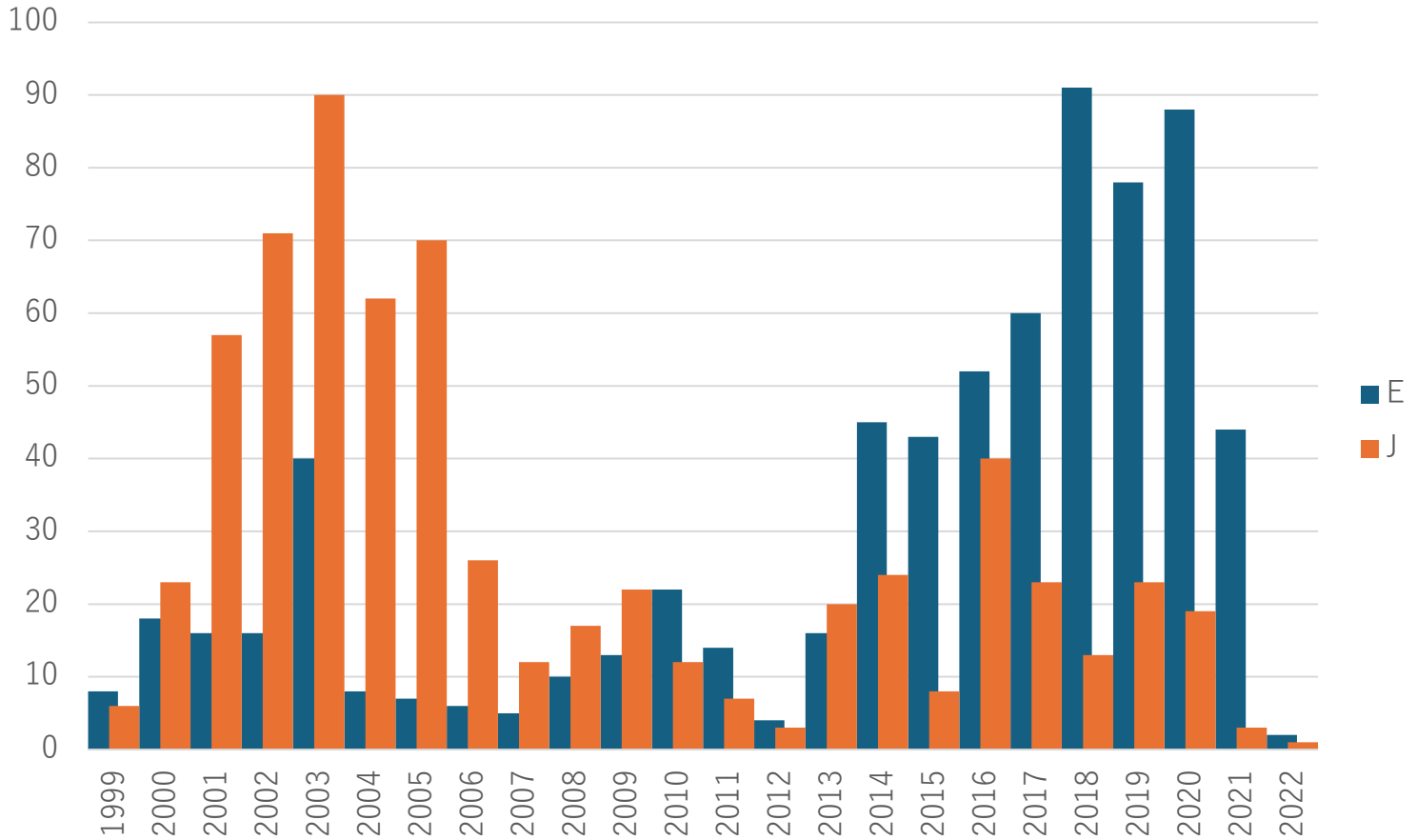
無償オンライン検索システム『中納言』のユーザー数（左軸）と年間検索数（右軸）

CSJを利用した研究の広がり：工学と言語



- 工学系（Engineering）と言語系(Language)に大別して引用件数を比較
- 第１波(1999-2010)と第２波(2012-2020)がある

CSJを利用した研究の広がり：執筆言語



- 英語（English）と 日本語(Japanese)
- 2010年代に入って英語論文が急増

現状の問題点と今後の課題

現状の問題点

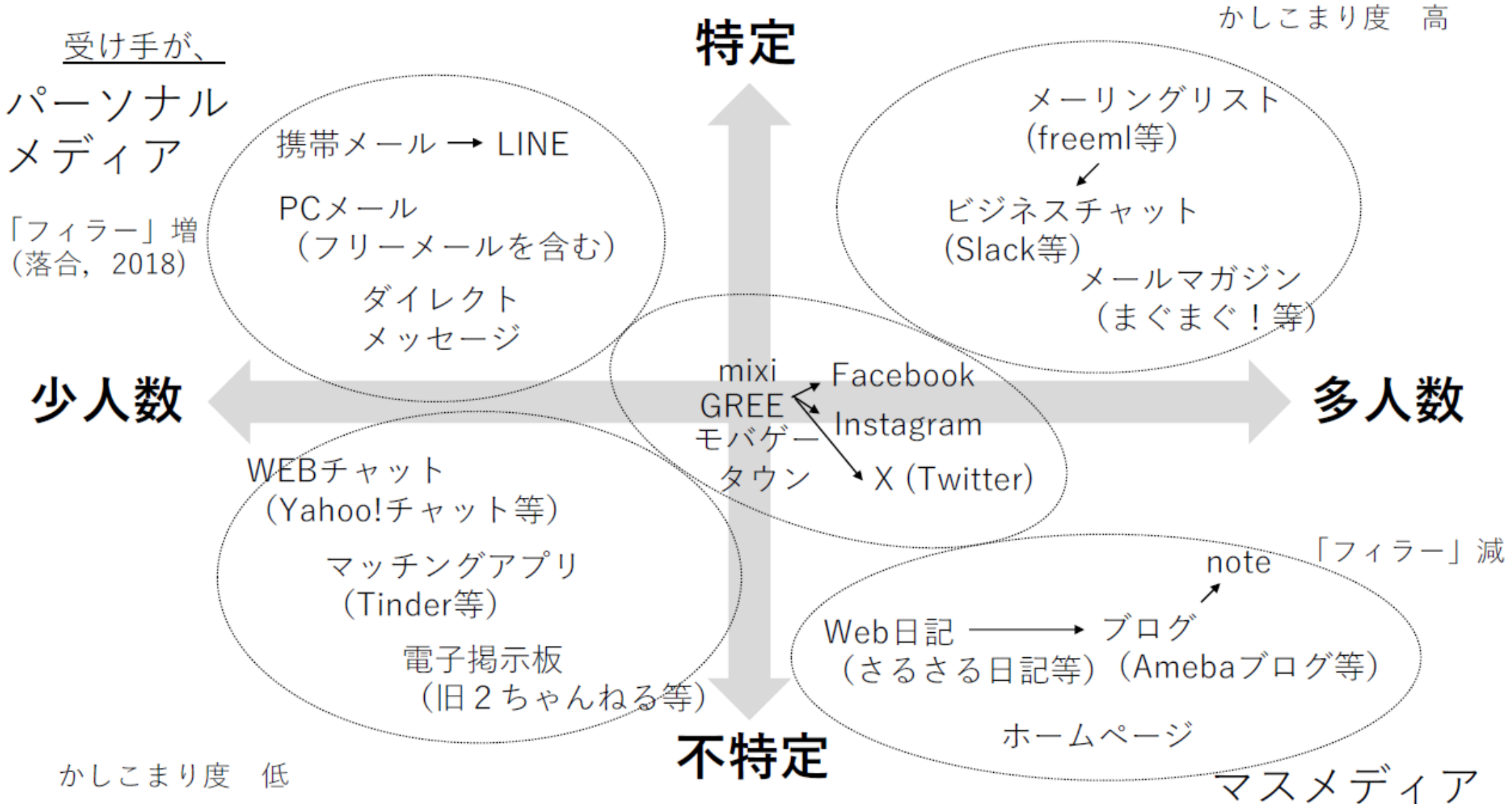
- 様々なコーパスが開発されているが小規模なものが多い
- 継続的な開発がおこなわれていない
- 素材はWebかBCCWJであることが多い（書き言葉の場合）
- FAIR原則が十分に満たされていない（特にinteroperability）
- 歴史的典籍は文学作品中心
- 権利処理はいまだに大きな問題
- まだまだ整備されていない／足りないレジスターが多い

構築が望まれるコーパス

• 在日・渡日外国人会話	多言語化社会への対応
• SNS (CMC)関係	現代のコミュニケーションの主体
• マンガ、アニメ、ゲーム	世界にとっての「日本」語
• 手話	「多言語」のひとつ
• 非典型話者音声言語	高齢者, ASD, 難聴者, L2 etc.
• 公共放送の音声・映像	共通語の成立基盤,「規範性」の要因
• 近世以降の文献	江戸以降は文献が非常に多い
• 主要コーパスのモニター化	BCCWJなどの経時的拡張
• AIが生成した日本語	「多言語」のひとつ, explainable AIの手掛り
• 各種センサーデータ	調音運動、視線運動など
• Plain language corpus	非母語話者に理解／生成が容易な日本語

国語研の同僚の見解も参照しました

さまざまなCMC　落合哉人(2024)の発表資料より引用



言語問題の俯瞰図
(Roadmap 2023申請書より)

社会の急激かつ多面的な変化

既存データ

- 国語研の既存言語資源群
- 人文機構DHによるオープンデータ
- 国会図書館によるオープンデータ
- 省庁・自治体等のオープンデータ
- 大学等のオープンデータ

etc.

既存データだけでは分析不能

サイバー化,
高齢化,
多言語化…

問題

問題

問題

問題

大変化期にある日本語

種々のコミュニケーション問題の発生

新規構築データ

- ①BCCWJ 拡張 2億語
- ②BCCWJ 通時拡張 1.4億語
- ③マンガ・アニメ 5000万語
- ④作文縦断 1500名×３テーマ×6年
- ⑤ASD話し言葉 60万語
- ⑥高齢者話し言葉 144万語
- ⑦L1L2接触場面話し言葉 100万語
- ⑧非典型話者調音運動 50時間
- ⑨ 日本の多言語状況調査１
- 日本の多言語状況調査２
- 日本の多言語状況調査３
- ⑩ 読み書き能力調査 簡易調査
- ⑪ 読み書き能力調査 詳細調査（第一次）

大規模データに対応でき、予測を可能にするE3P言語学の開拓

【音声科学】
先進的な機器を用いた音声生成モデルの構築

【認知科学】
言語の受容過程のデータ収集及び分析

【統計モデリング】
言語情報・センシングデータの統計的分析

【言語学】

言語使用のデータ収集及び分析
【コーパス】

言語理論・言語事実の共時的・通時的検証
【言語理論】

言語実験棟の新設（MRI装置等）

期待されるimpact：
「新しい言語観」

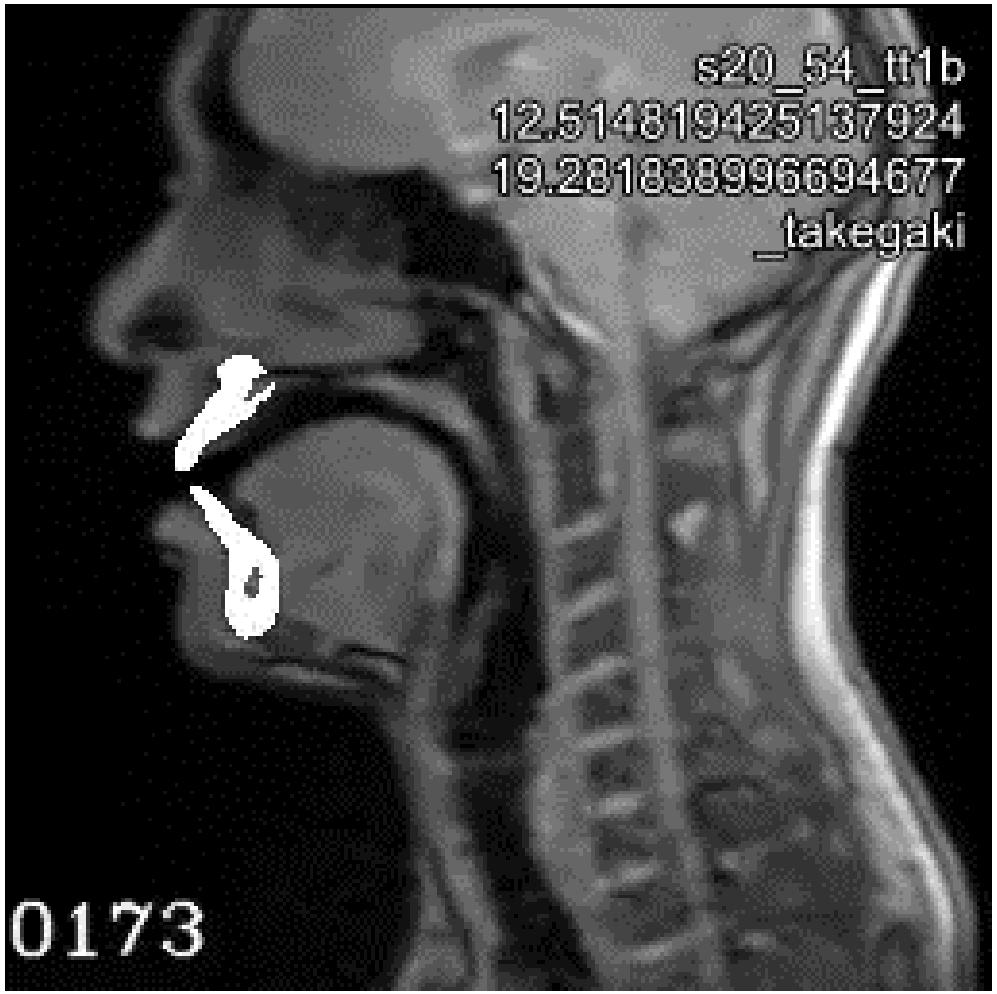
日本語の変化要因　(前川2024)

- インターネットの影響
  - 「打ち言葉」の圧倒的普及
  - メール, SNS等の媒体の定着
- AIの影響
  - 機械翻訳された日本語との接触
- 日本社会の多言語化
  - 定住外国人380万人超

ウェブのマンガサイトの広告から

⇒日本国内の言語状況が全く把握できていない
（どれだけの言語をどれだけの人がどういう目的で使っているか, etc.）

センサーデータの例：リアルタイムMRI



Cf. Maekawa (In press)

まとめ

- 過去四半世紀で日本語コーパスの整備は顕著に進み、言語研究にしめるコーパスの位置は確立されたと言える
- しかし未だ路半ばであり、今後データ整備が望まれる多くのレジスターが残されている
- とりわけ、将来の日本語に強い影響を及ぼすと考えられるCMC関係や日本社会の多言語化に関する部分が弱い
- 一方で情報処理領域での需要も著しく増大している
- 国費を組織的に投入するのであれば、整備の順序について超領域的な熟議が必要
- また諸外国における政策の比較研究も必要

文献／URL

淺原正幸ほか6名「Universal Dependencies日本語コーパス」自然言語処理 28 (1), 2019.

石川慎一郎「英語学習者コーパス研究の現状と課題」IEICE Fundamentals Review, 12 (4), 280-289.

落合哉人「日本語環境における CMC の分類と X (Twitter) の特徴に関して」第272回NINJALサロン, 2024.06.04.

前川喜久雄「コーパスの存在意義」前川編『コーパス入門』講座日本語コーパス第1巻, 朝倉書店, 2013.

前川喜久雄「仮想講義 言語資源学入門」（特集：データが変えることばの研究と教育）日本語学,35 (13), pp. 2-11, 2016.

前川喜久雄「これからの日本語研究と国立国語研究所：E3P-Linguisticsをめざして」言語処理学会第30回年次大会(NLP2024)特別招待講演, 2024.03.14. https://www.youtube.com/watch?v=CMoO8x5wtEQ

Maekawa, K. “Real-time MRI articulatory movement database and its application to articulatory phonetics”, *Acoustical Science & Technology*, In press.