NINJAL
National Institute for Japanese Language and Linguistics

国立国語研究所の言語資源とオープンデータ・オープンサイエンス

小木曽智信

2019/10/24 NII

大学共同利用機関法人 人間文化研究機構
国立国語研究所
National Institute for Japanese Language and Linguistics

自己紹介

- 小木曽 智信（おぎそ としのぶ） togiso@ninjal.ac.jp
  - 人間文化研究機構 国立国語研究所 言語変化研究領域代表 （コーパス開発センター兼任）
  - 共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」プロジェクトリーダー（2016年～）
  - 専門は**日本語学、自然言語処理**
  - 『太陽コーパス』『現代日本語書き言葉均衡コーパス』『日本語歴史コーパス』など書き言葉のコーパス構築に携わる

国語研と
オープンサイエンス・オープンデータ

NINJAL
National Institute for Japanese Language and Linguistics

オープンサイエンス・オープンデータを基盤に

- オープンなデータ
  - 調査収集したデータは公開を原則とする
  - 所定の手続きで誰でもアクセス可能に
    ※必ずしも無償・無制限であることまでは意味しない

- 研究方法もオープンに、検証可能に
  - 実験など主観を排した方法で
  - 研究に用いた中間データやスクリプト類もできるだけ公開

第4期の国語研の基本方針(案)より

コーパスとアーカイブを核に

- これまでに高い評価を受け、今期の「多様な言語資源に基づく総合的日本語研究の開拓」でも多数構築中のコーパスを中心とする言語資源を更新・拡張
- 危機言語データ等のアーカイブに取り組み、オープンサイエンスを支える

第4期の国語研の基本方針(案)より

次期、オープンサイエンスを謳う理由

現在はまだできていないから

- データの囲い込み（個人作成・収録データ）
- エビデンスが見えない文法性判断に基づく議論
- 論文の元となった（中間）データの非公開・・・

それでも、それに向けた取り組みは始めている

国語研のコーパスとオープンデータ

国立国語研究所のコーパス

コーパスとは：言語を分析するための基礎資料として，書き言葉や話し言葉の資料を体系的に収集し，研究用の情報を付与した大規模なデータベース

- 『日本語話し言葉コーパス』(CSJ) 2004年
- 『現代日本語書き言葉均衡コーパス』(BCCWJ) 2011年
- 『日本語歴史コーパス』(CHJ)　2013年～
- 『国語研日本語ウェブコーパス』 2016年
- 『多言語母語の日本語学習者横断コーパス』2016年～

※現在は日常会話、方言などのコーパスも構築中

コーパス
開発センター
Center for Corpus Development, NINJAL

コーパス開発センター

NINJAL
National Institute for Japanese Language and Linguistics

https://pj.ninjal.ac.jp/corpus_center/

研究インフラとしてのコーパス

- 「現代日本語書き言葉均衡コーパス」
  - 登録ユーザ数：約20,000人
  - 年間クエリ数：約50万件/年
  - 利用した論文数：約70本/年
- 「日本語歴史コーパス」
  - 登録ユーザ数：約10,000人
  - 年間クエリ数：約26万件/年
  - 利用した論文数：約50本/年（※予稿集含む）

コーパス検索アプリケーション「中納言」

- 人文系の研究者に利用しやすい形で提供
  - オンライン
  - 無料（要登録）

NINJAL
National Institute for Japanese Language and Linguistics

コーパス構築のコスト

「現代日本語書き言葉均衡コーパス」(BCCWJ)
2006～2010年(研究代表者:前川喜久雄)

- 約1億語の現代語書き言葉のコーパス
  - 書籍等からのバランスをとったサンプリング
  - 紙の文献の電子化
  - 単語情報の付与
- 予算:
  - 科研費 特定領域研究(総額 約8億円)
    ＋国語研運営費交付金

コーパス構築のコスト2

「日本語歴史コーパス」(CHJ)
2013年～現在(研究代表者:小木曽)

- 奈良時代から明治・大正時代までの日本語
  - 全本文に単語情報の付与
  - 外部の画像等にリンク

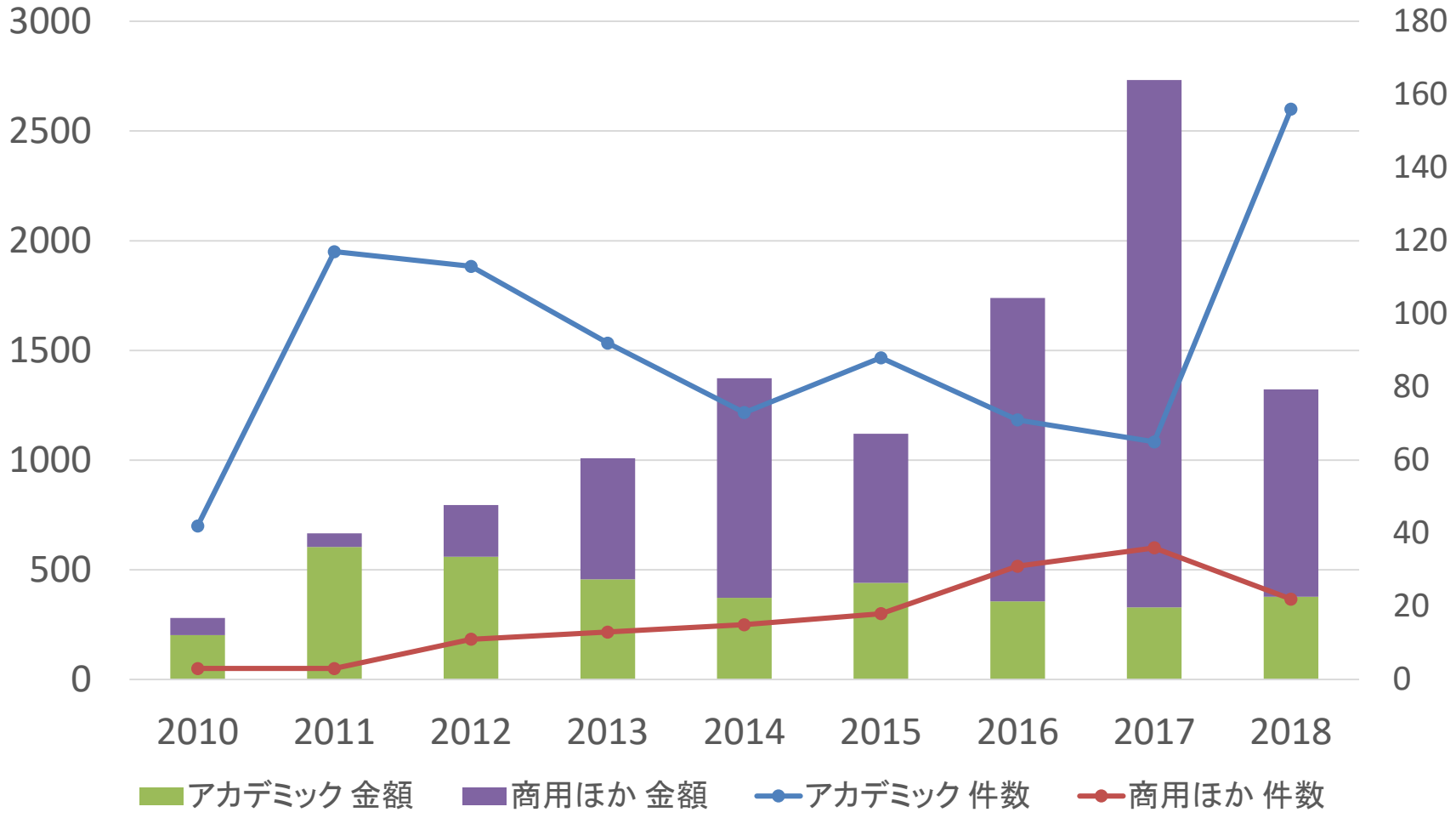
  ※2013年以前に構築されたデータを継承して含む
- 予算:
  - 国語研運営費交付金より 年間3000万弱
    ＋科研費 基盤(A) 年間1000万弱

コーパス公開のコスト

- サーバー代等：概ね年間1000万円程度
  （＋人件費、電気代 etc.）

- 新しいコーパスを構築する際は、その新規性をもって予算獲得ができるが、インフラとして定着したコーパスの公開・維持費は外部資金では困難

コーパスからの収益


3000
2500
2000
1500
1000
500
0
180
160
140
120
100
80
60
40
20
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
アカデミック 金額
商用ほか 金額
アカデミック 件数
商用ほか 件数

コーパスとオープンデータ

- コーパスは研究の副産物ではなく、その構築自体が目的でもある
- コーパスは他者の著作物ではなく、構築者自らの（編集）著作物である
- コーパスは自己収入を生み出す経済的価値を持つ

- ➢ 完全なオープン化は困難
- ➢ 所定の手続きで誰でもアクセス可能にするが、※必ずしも無償・無制限であることまでは意味しない

コーパスとオープンデータ

- コーパス本体はオープンデータ化は難しいが、コーパスに対するアノテーションは公開できる
  - アノテーション：追加情報の付与
    - コーパス中の単語や文、発話・・・に
    - 意味情報、統語情報、メタ情報・・・を付与
- コーパスに依拠したオープンサイエンスの可能性

『日本語歴史コーパス』を例に

コーパスとオープンサイエンス

NINJAL
National Institute for Japanese Language and Linguistics

- これまでの日本語研究では、コーパスを利用した場合でも、コーパス中の用例を分類したり分析をおこなったりした資料は個人のもとで管理され公開されることはなかった。研究が論文にまとめられると、その基礎となったデータは再利用されず死蔵され、場合によっては破棄される場合も少なくない。

- しかし、コーパスに依拠したデジタルデータは、インターネット上で共有することは比較的容易である。基礎データの共有が可能になれば、当該の研究が検証可能になるだけでなく、その成果を再利用して他の研究者が新たな研究を行うことも可能になる。コーパスを対象とした研究データの多くは、コーパスに対するアノテーションとしてまとめられる。

オープンサイエンスに向けたコーパス検索ツール側の仕組み

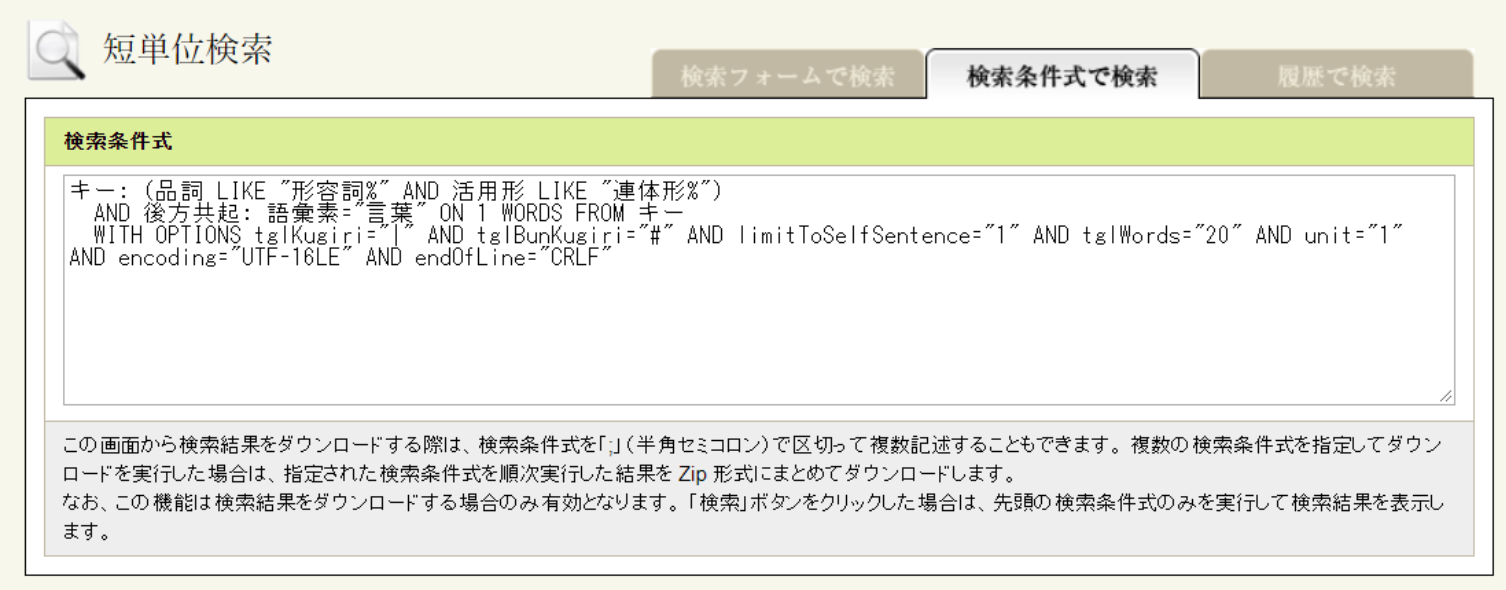
1. 検索条件式とその共有
2. 用例へのパーマリンク
3. 用例のユニークIDとアノテーション
4. アノテーションデータの共有
5. アノテーション共有環境の構築

1. 検索条件式

- 「中納言」で“「言葉」という語の直前に来る形容詞の連体形を検索”

1. 検索条件式

- 「中納言」で“「言葉」という語の直前に来る形容詞の連体形を検索”



- 研究の再現性、用例の共有

1. 検索条件式

- 「検索条件式」を使うことで、中納言ユーザーなら誰でも、同じ検索を行うことができる

```
キー: (品詞 LIKE "形容詞%" AND 活用形 LIKE "連体形%") AND 後方共起: 語彙素 = "言葉" ON 1 WORDS FROM キーWITH OPTIONS unit="1“
```

- 研究の再現性のために論文などで使用した検索条件式を明記するとよい

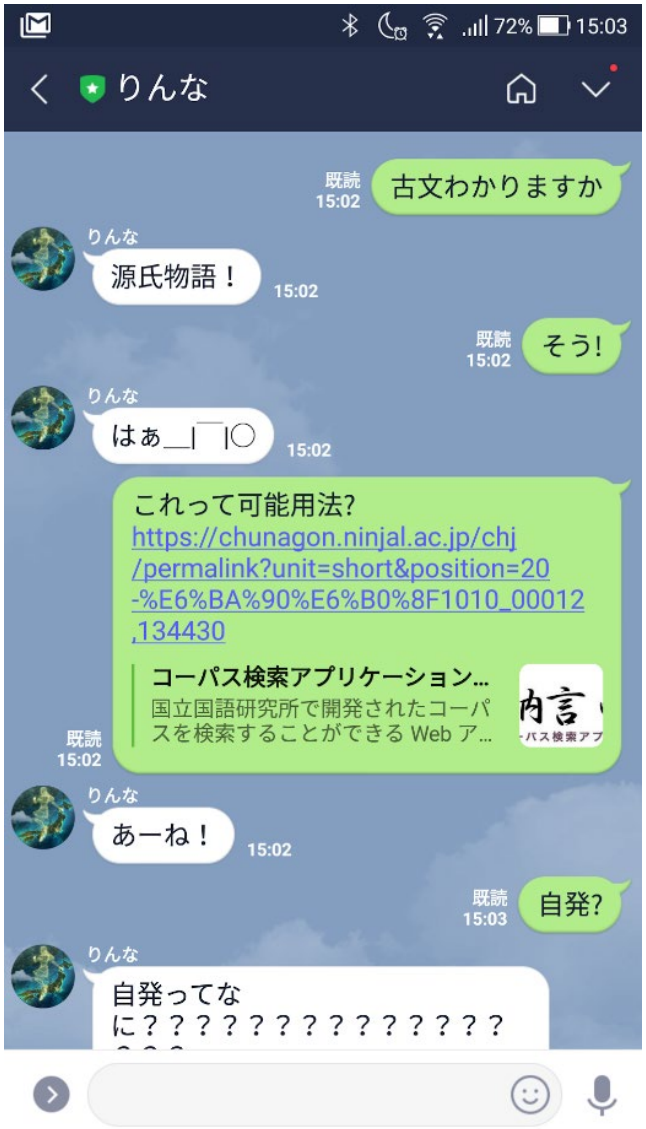
➢ 研究データの共有、検証可能性

2. 用例へのパーマリンク

- 位置情報をもとにした中納言上の用例へのリンク
  https://chunagon.ninjal.ac.jp/chj/permalink?unit=short&position=20-源氏1010_00012,134430

  - 論文や資料のPDFに用例へのリンクを埋め込み！
  - QRコードにも！　SNSでも！

2. 用例へのパーマリンク



NINJAL
National Institute for Japanese Language and Linguistics

3. 用例のユニークID

『日本語歴史コーパス』の位置情報

- サンプルIDと開始位置（＝先頭からの何文字目か）
  - 20-源氏1010_00012, 134430

- 位置情報で用例が一意に定まる＝用例のマイナンバー

サンプルID	開始位置	連番	コア	前文脈	キー	後文脈	語形	品詞
20-源氏1010_00012	134430	77920	1	そ|つと|さぶらひ|ける|。# |前栽|の|花|いろいろ|咲き|乱れ|、|おもしろき|夕暮|に|、|海|見|やら|	るる	|廊|に|出で|たまひ|て|、|たたずみ|たまふ|御|さま|の|ゆゆしう|きよら|なる|こと|、|所がら|は|まして|に	ル	助動詞
				うまつるをうれしきことにて、四五人ばかりぞつとさぶらひける。# 前栽の花いろいろ咲き乱れ、おもしろき夕暮に、海見やら	るる	廊に出でたまひて、たたずみたまふ御さまのゆゆしうきよらなるこ		

3.用例のユニークID

用例集としての位置情報

- 用例の位置情報を集めて並べればそれだけで貴重な用例集になるかも！
  - あとで再利用できる、検証できる
  - 「中納言」経由で現代語訳や原文画像が参照できる

位置情報を用いたアノテーション

- 位置情報に、用例の情報を追加すれば立派なアノテーションデータ！

  20-源氏1010_00012, 134430, 自発

4. アノテーションデータの共有

- 小木曽(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 予稿集

サンプル ID	開始位置	連番	本文種別	語形	用法
20-伊勢0920_00001	1730	1070	歌	ル	可能
20-伊勢0920_00001	29880	19170		ラル	受身
20-伊勢0920_00001	68940	44330		ラル	受身
20-伊勢0920_00001	73050	47020	歌	ル	自発

4.アノテーションデータの共有

- https://researchmap.jp/mu1dor8so-12361/

4.アノテーションデータの共有

- データの配布と引用
  - 小木曽(2019)「CHJ 中古「る」「らる」用法分類アノテーションデータ」 https://researchmap.jp/mu1dor8so-12361/

- オープンデータとして「公開」しよう！
- アノテーションを「引用」しよう!!

  → コーパスをみんなで「育て」よう!!!

4.アノテーションデータの共有

「……ともに「コーパスを育てていく」作業に参加してほしい。こうして作られ、公開されたアノテーションデータは他の研究者に役立つ共有の財産となるものであるから、このようなデータの作成・公開が、研究の業績として正当な評価を得られるようになることも願っている。」

小木曽(2019)

5.アノテーション共有環境の構築

- ユーザーがもっと容易にアノテーションを付与し、共有し、引用し、評価しできる環境をアプリケーション上で実現したい・・・
- ユーザーが多い「中納言」の機能拡張で・・・

科研・挑戦的研究(開拓) 19H05477

「日本語コーパスに対する情報付与を核としたオープンサイエンス推進環境の構築」

- 研究代表者:小木曽智信
- 2019～2021年度
- 25,740千円


KAKEN 研究課題をさがす 研究者をさがす KAKEN

前のページに戻る

日本語コーパスに対する情報付与を核としたオープンサイエンス推進環境の構築

研究課題/領域番号	19H05477
研究種目	挑戦的研究(開拓)
配分区分	補助金
審査区分	中区分2:文学、言語学およびその関連分野
研究機関	大学共同利用機関法人人間文化研究機構国立国語研究所
研究代表者	小木曽 智信　大学共同利用機関法人人間文化研究機構国立国語研究所, 言語変化研究領域, 教授 (20337489)
研究分担者	河内 昭浩　群馬大学, 教育学部, 准教授 (10625172) 橋本 雄太　国立歴史民俗博物館, 大学共同利用機関等の部局等, 助教 (10802712) 永崎 研宣　一般財団法人人文情報学研究所, 人文情報学研究部門, 主席研究員 (30343429) 馮野 知暁　東京大学, 大学院人文社会系研究科(文学部), 助教 (30751515) 海野 圭介　国文学研究資料館, 研究部, 教授 (80346155) 後藤 真　国立歴史民俗博物館, 大学共同利用機関等の部局等, 准教授 (90507138)
研究期間 (年度)	2019-06-28 – 2022-03-31
研究課題ステータス	交付 (2019年度)
配分額 *注記	25,740千円 (直接経費: 19,800千円、間接経費: 5,940千円) 2019年度: 8,970千円 (直接経費: 6,900千円、間接経費: 2,070千円)
キーワード	コーパス / アノテーション / オープンデータ / 人文情報学 / 日本語学

NINJAL
National Institute for Japanese Language and Linguistics

科研・挑戦的研究(開拓) 19H05477

- 「中納言」に、利用者が新たな情報を任意の場所に付加するアノテーション機能を追加
- 付加した情報を他のユーザーと共有できるシステムを構築
- 公開されたアノテーションデータを引用・評価できる仕組みを提供

➢ この環境の構築とアノテーションの実践により、研究データの公開と共有を促し、書き言葉テキストを中心としたオープンサイエンスの基盤とすることを目指す。

まとめ

- 国語研のコーパスとその検索環境はすでに日本語研究のインフラとして機能している。
- コーパスの完全なオープンデータ化は困難だが、コーパスとアノテーションの共有を通してオープンサイエンスを推進する基盤となしうる。
- 検索環境を拡張し、言語研究のオープンサイエンス推進環境の整備をすすめ、今後につなげたい。