NINJAL
National Institute for Japanese Language and Linguistics

国立国語研究所における言語資源開発（これまでとこれから）

前川　喜久雄

国立国語研究所

言語資源研究系・コーパス開発センター

言語処理学会創立20周年おめでとうございます

講演の趣旨

- 言語処理学会の創立（1994年）以前に国語研が実施した言語資源よりの調査活動を紹介する
- 国語研による（今日的な意味での）言語資源開発の現状を紹介する
- 言語資源開発のこれからの課題を展望する

国立国語研究所

- 1948年　創立。戦後初の国立試験研究所
- 1968年　文化庁設置に伴い文化庁へ
- 2001年　独立行政法人へ移行
- 2009年　大学共同利用機関法人（人間文化研究機構）へ移管

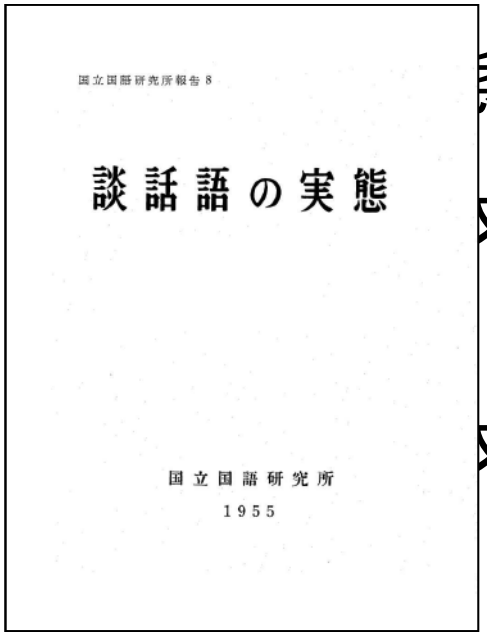
初期の国語研における「言語資源」関連研究

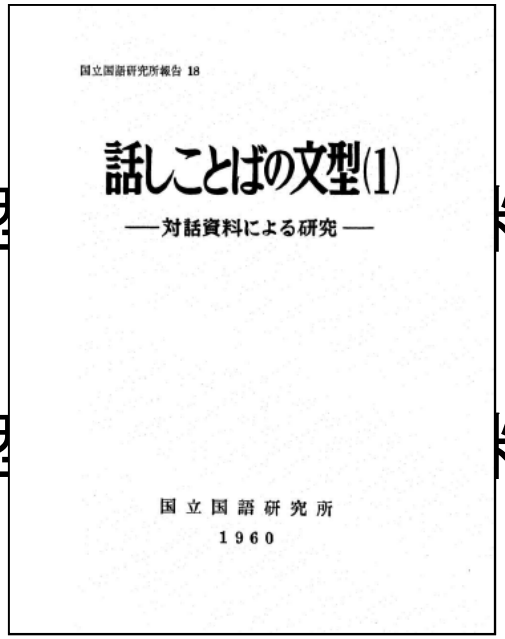
- 話しことば研究
  - 当時未開拓であった現代語研究の方法論構築の一環として、話し言葉の研究法を開拓した。大量かつ多様な話し言葉のデータを収集して分析した
- 語彙調査
  - やはり現代語の書きことば研究の基礎として、推計学に基づく語彙調査を実施した。後に国語研の十八番となり、ながく継続された

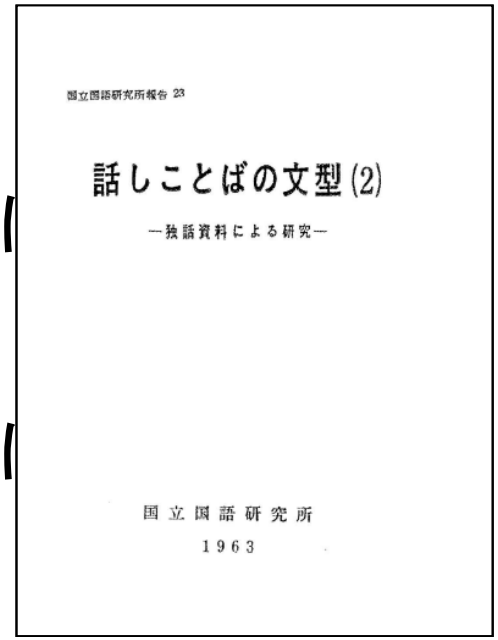
話しことば研究

話しことば研究の成果：3冊の報告書

- 『
- 文
- 文







『談話語の実態』(1955)におけるデータ収集：
オープンリール80巻

Reel No.	略称	録音状況の可否	地区 下町	地区 山の手	地区 周辺	場所 家庭	場所 近所	場所 学校	場所 職場	場所 公共	性 男	性 女	年齢 男 若	年齢 男 壮	年齢 女 若	年齢 女 壮	教養 職	教養 専	相手 1人	相手 2－4人	相手 5－7人	相手 8人以上	相手 未知	相手 既知
3	1 夫婦	可																	×				×	
7	T家雑談	可															×	×		×			×	
67	N家座談	可															×			×			×	×
86	トタン屋	可															×	×		×			×	
76	じいさんばあさん	可															×			×			×	×
93	魚屋雑談	可															×	×		×				×
97	U氏談	可																		×			×	×
61	学生座談1	可																			×			×
66	井戸端会議	可															×	×			×			×
98	友の会	可															×	×		×				×
2	女子学生	可																			×			×

データ収集基準

- 地域: 山の手～下町～郊外
- 場所: 自宅、近所、学校、職場、公民館、等
- 性別: 男女のすべての組み合わせ（4通り）
- 年齢: 若年と中年のすべての組み合わせ（4通り）
- 学歴: 低、中、高
- 話者数: 独話と様々な会話（2～8名）

話し言葉について「均衡した」データを作成しようとした試みとしては世界最初と思われる。

肩掛け録音器による
話しことばの収録

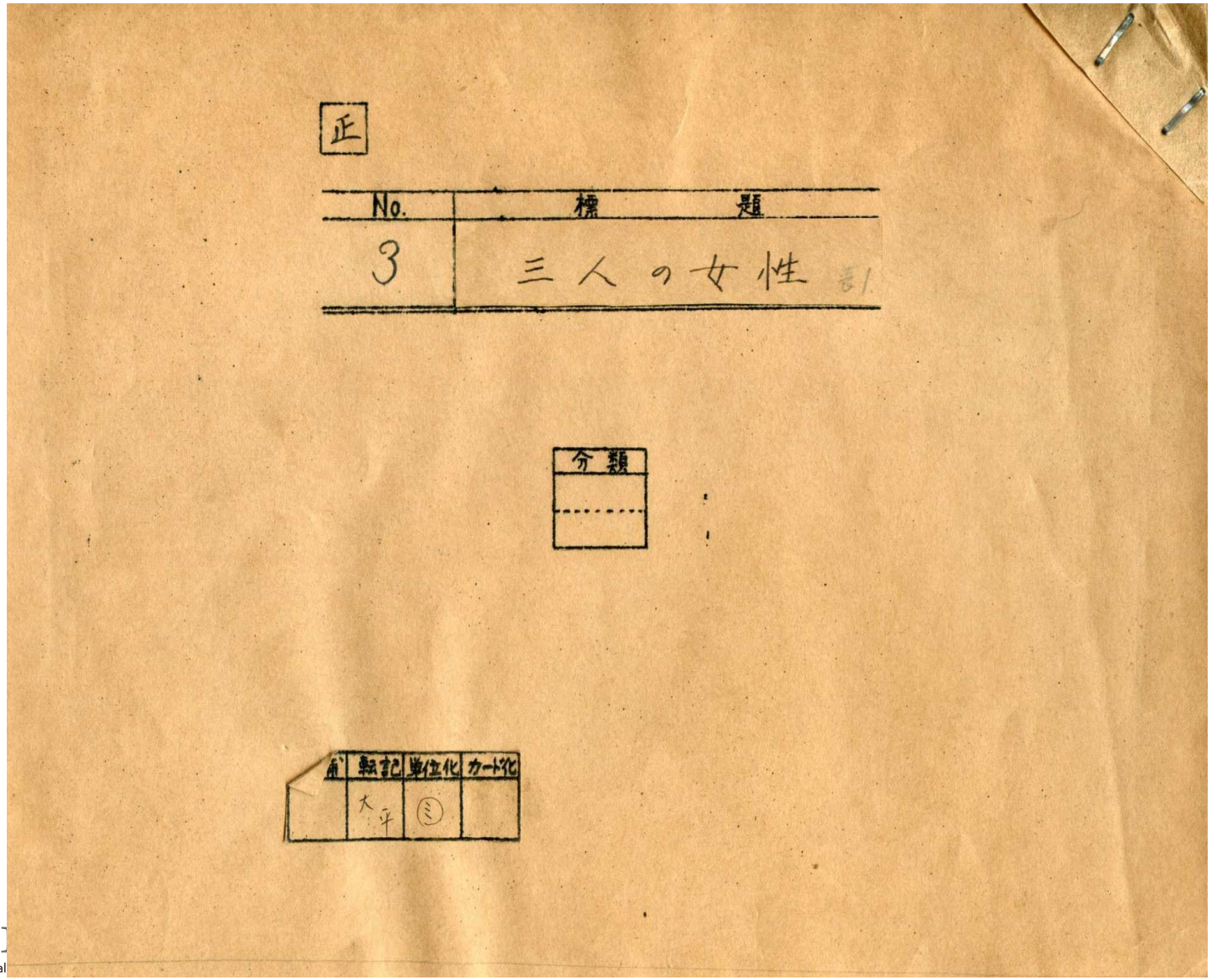
肩掛け録音器

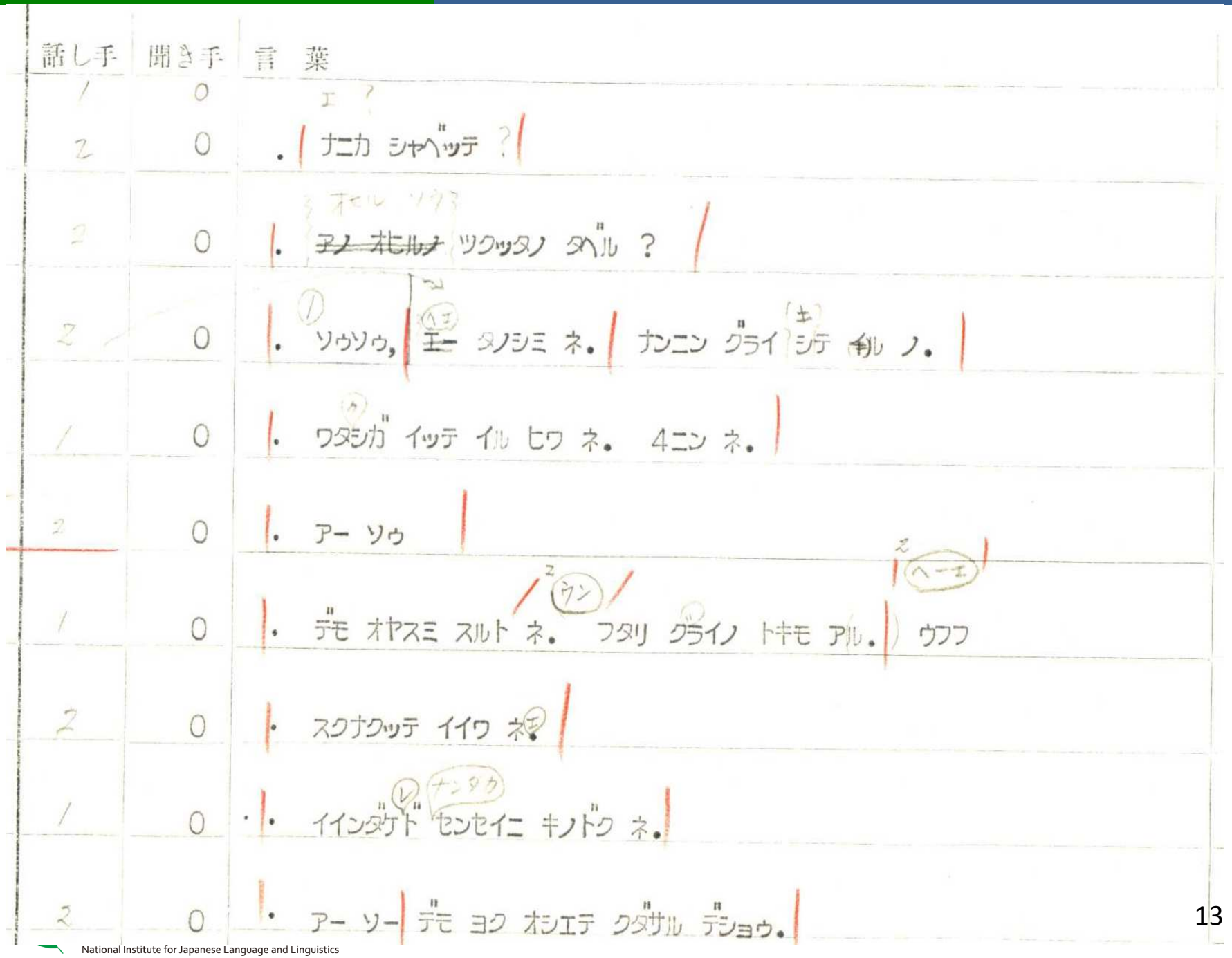
残存するテープ

- 職場での電話会話
  - 録音年代不詳。おそらく1950年代末。
  - 国語研の研究者同士（男性、ともに1911年生）

- 若い女性3名（すべて20代）の雑談
  - 1957年2月録音
  - 『話しことばの文型(2)』で分析されたサンプル

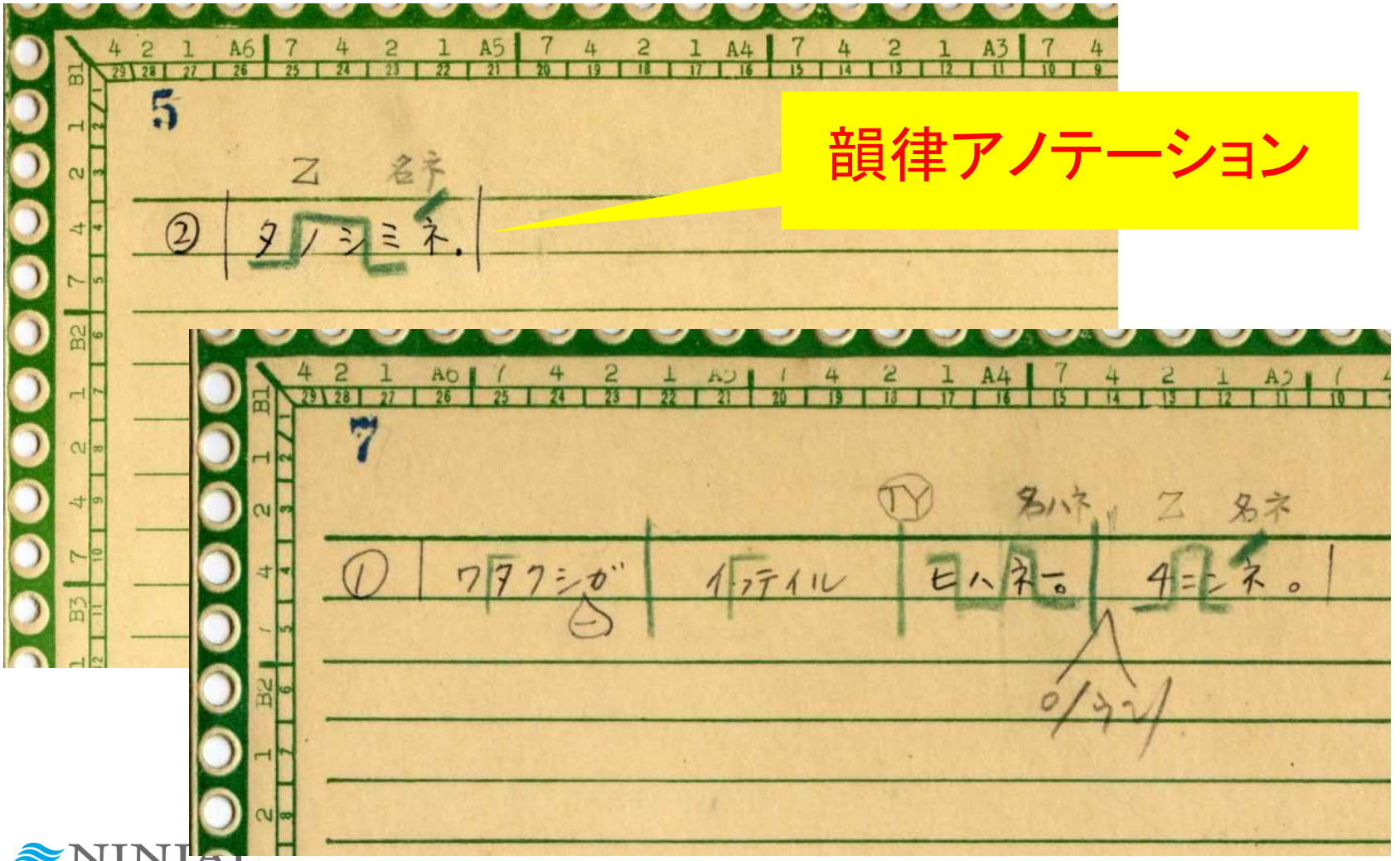
『談話語の実態(1)』のための転記テキスト





話し手	聞き手	言葉
1	0	エ？
2	0	ナニカ シャベッテ ？
2	0	~~アノ 北ルノ~~ ツクッタノ タベル ？
2	0	ソウソウ, エー タノシミ ネ。 ナンニン グライ シテ イル ノ。
1	0	ワタシガ イッテ イル ヒワ ネ。 4ニン ネ。
2	0	アー ソウ
1	0	デモ オヤスミ スルト ネ。 フタリ グライノ トキモ アル。 ウフフ
2	0	スクナクッテ イイワ ネ。
1	0	イインダケド センセイニ キノドク ネ。
2	0	アー ソー デモ ヨク オシエテ クダサル デショウ。

機械式ソーティングカード



停滞と再生

- 『話しことばの文型(2)』(1963)で、一連の話しことば研究は終了。詳しい経緯は不明
- 収集されたデータも一部を除いて散逸。非公開
- 話しことば研究は60年代後半から停滞期に突入。国語研以外の実施主体は育っていなかったので、結局、日本全体が停滞
- この方面の研究が再開されるのは、30年後の『日本語話し言葉コーパス』構築プロジェクト(1999~2003年)

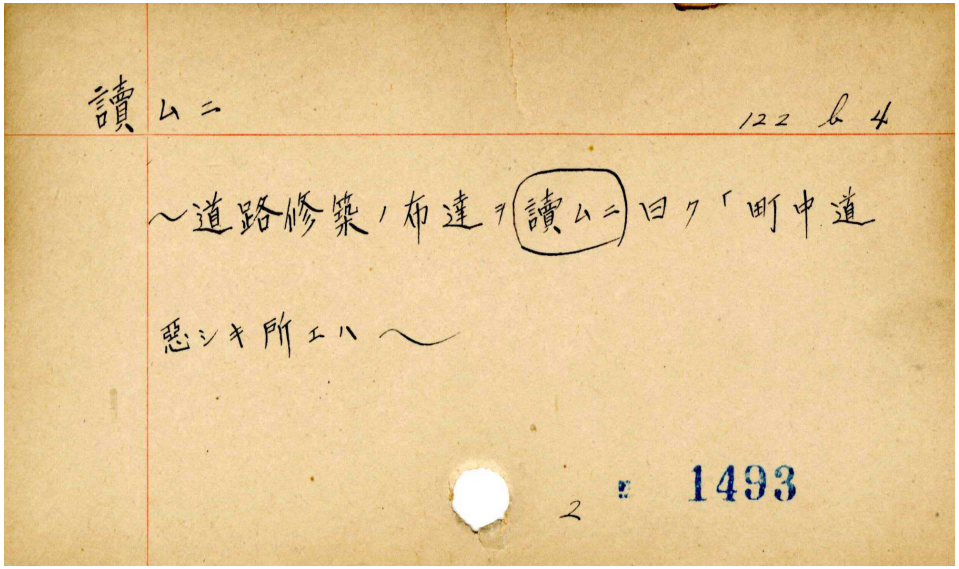
語彙調査

語彙調査の目的

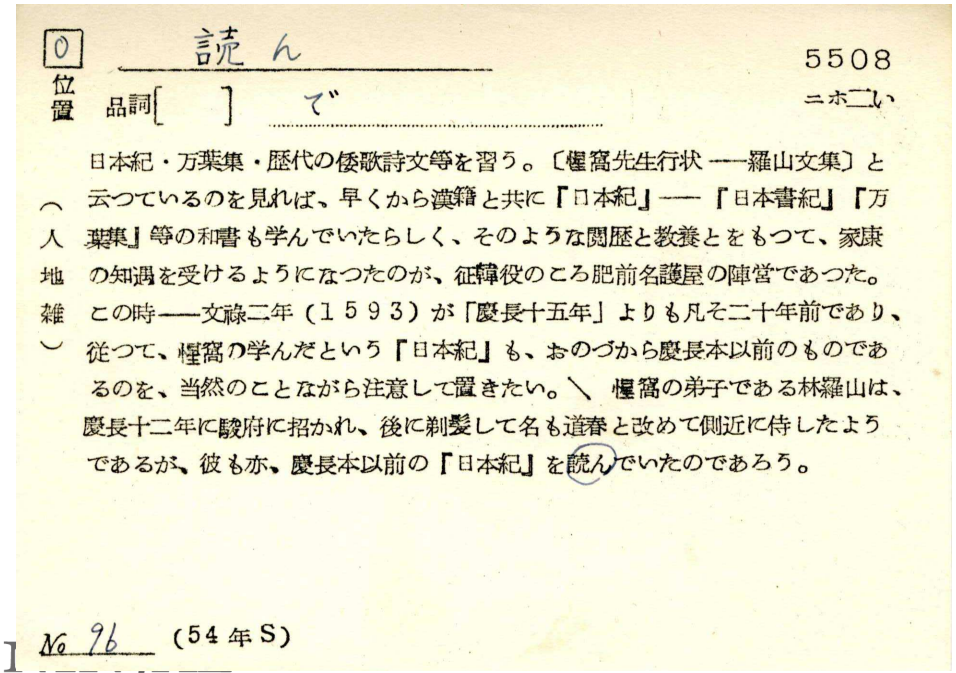
- 日本語の使用実態を記述・把握
  - 1950年代まで、現代語は国語学の研究対象外
- 資料による記述範囲の明確化
  - 行動主義的研究
- 基本語彙，生活語彙を重視
  - 戦後の「国語合理化」「言語生活研究」
- 国語国字問題解決の参考資料
  - 明治以来の文科省の悲願

国立国語研究所の語彙調査

調査資料（資料の期間）	調査方法	延べ語数	異なり語数	調査単位	報告書出版年
①新聞1か月（1949年6月）	全数調査	24万	1.5万	β’	1952
②婦人雑誌（1950年）	標本調査	15万	2.7万	α	1953
③総合雑誌（1953～54年）	標本調査	23万	2.3万	β	1957–58
④郵便報知（明治10年11月）	標本調査	10万	2.8万	文節	
⑤雑誌九十種（1956年）	標本調査	53万	4.0万	β	1962–64
⑥新聞3紙（1966年）	標本調査	300万	21.3万	短	1970–73
		200万	ー	長	
⑦高校教科書（1974年）	全数調査	59万	1.6万	W	1983–84
		45万	4.1万	M	
⑧中学校教科書（1980年）	全数調査	25万	0.8万	W	1986–87
		20万	1.8万	M	
⑨ﾃﾚﾋﾞ放送(1989年4～6月)	標本調査	14万	2.6万	長’	1995–99
⑩雑誌70誌（1994年）	標本調査	105万	4.8万	β	2005

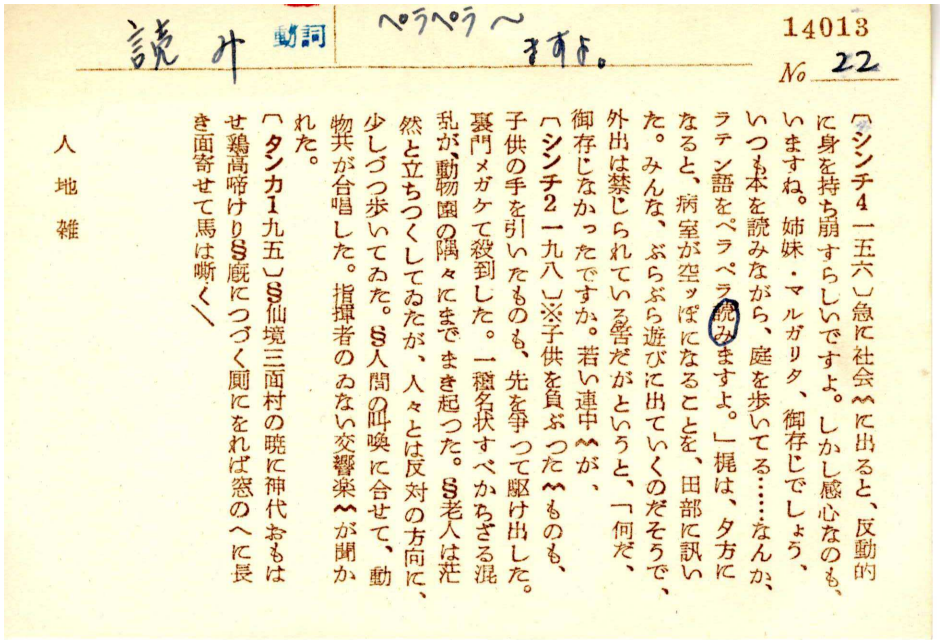


「新聞1か月」調査(1949)に用いられたと思われる手書き用例カード



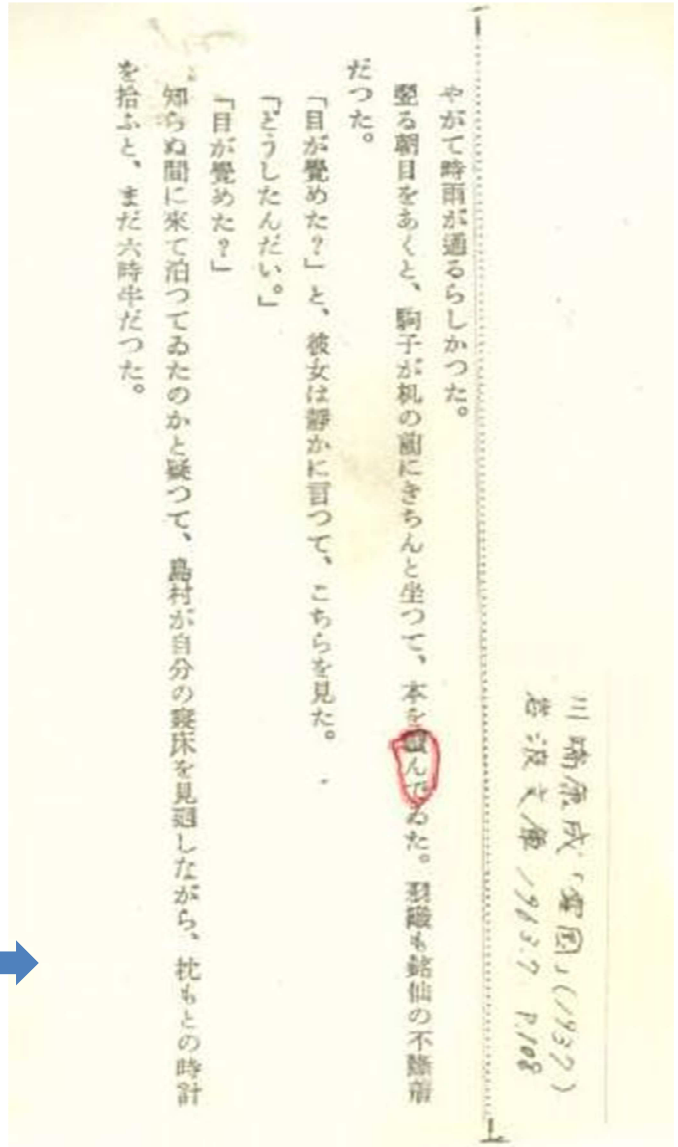
総合雑誌調査(1954)に用いられた和文タイプライタで作成した用例カード

資料提供：宮島達夫氏

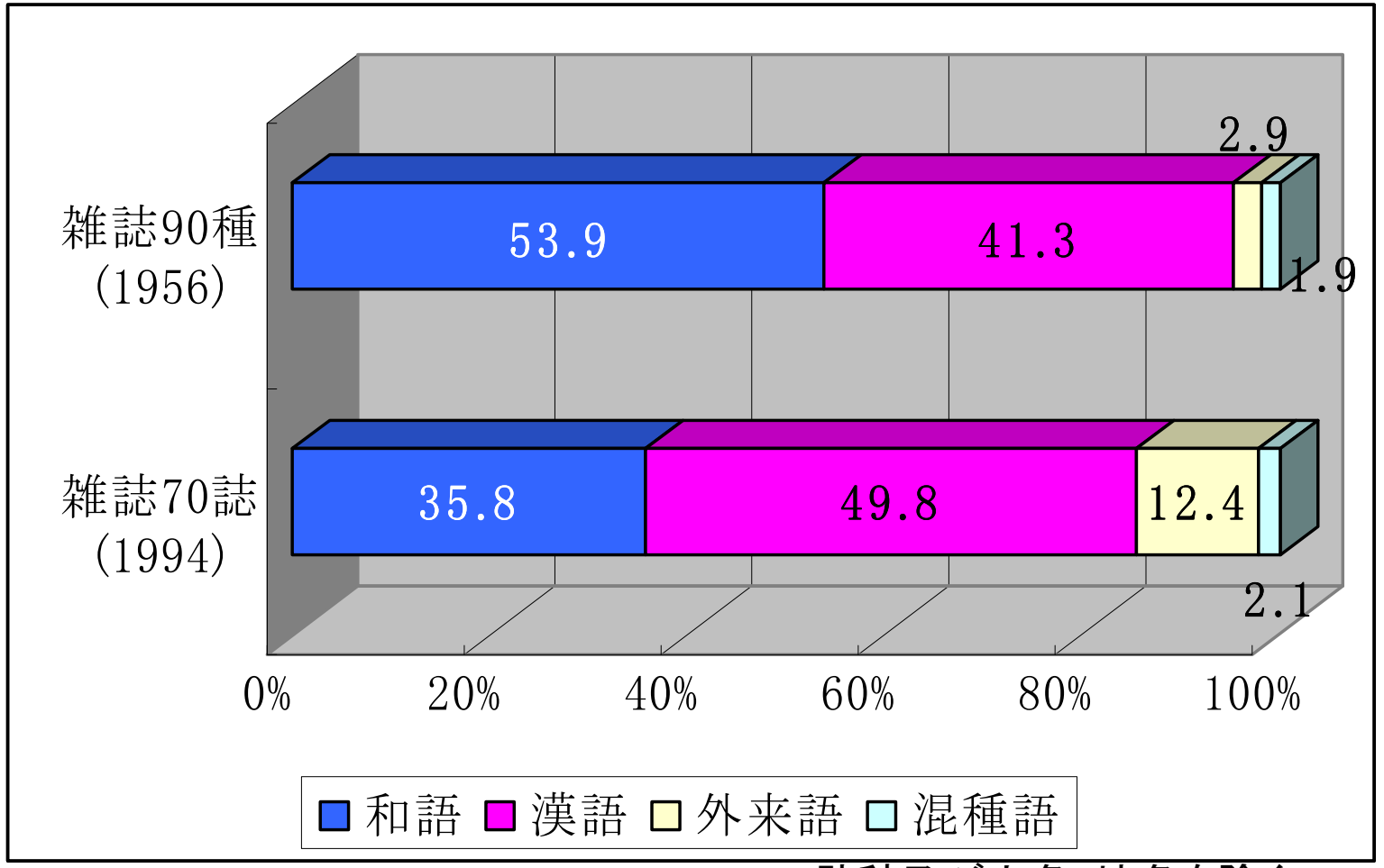


やはり「総合雑誌」調査の用例カード。
縦書き

1970年前後になってゼロックスが利用可能に（革命的に便利！）

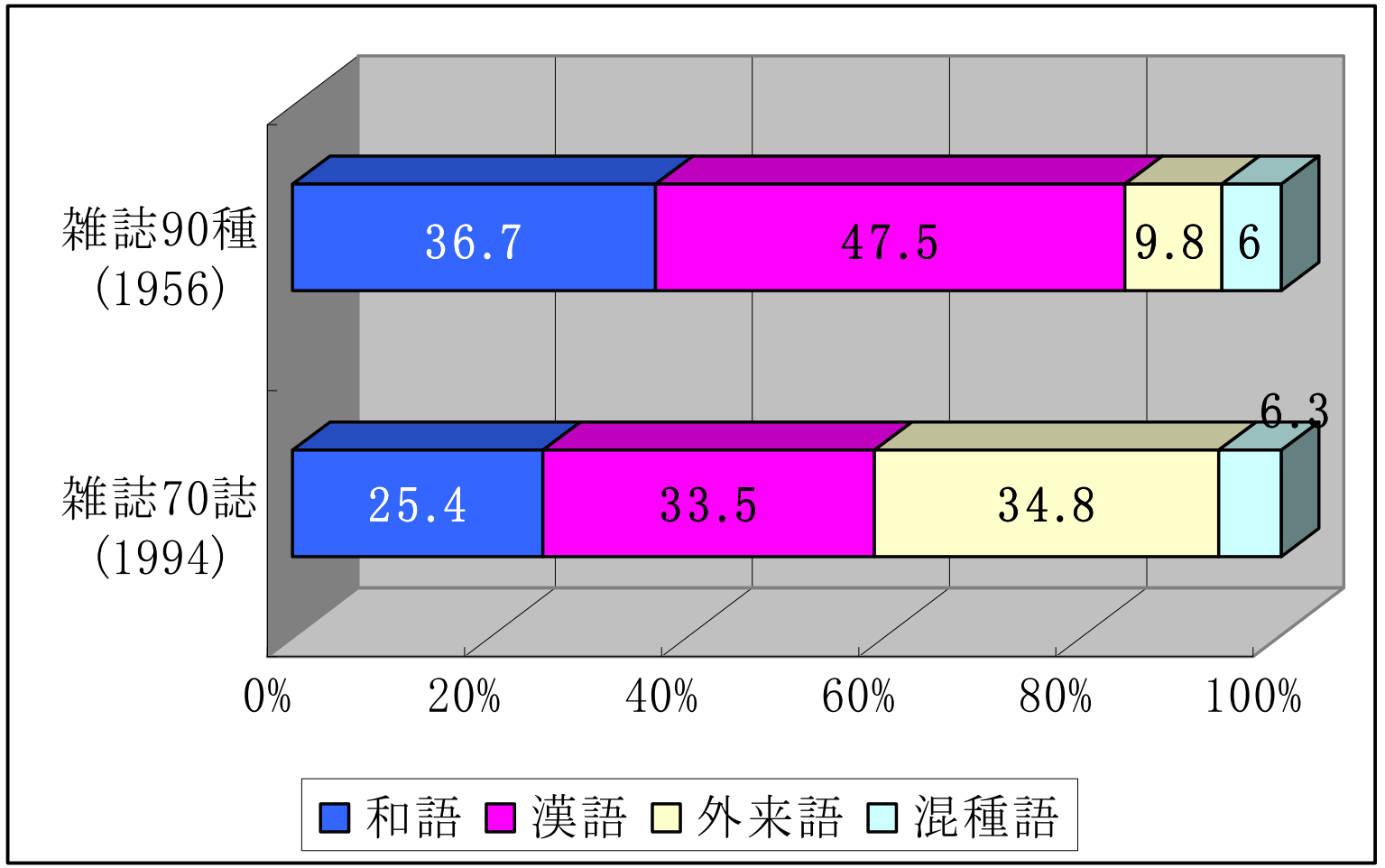


語種の構成（延べ語数）



助辞及び人名・地名を除く

語種の構成（異なり語数）



助辞及び人名・地名を除く

国語研語彙調査の問題と限界

- 調査単位の不統一
  - β単位、M単位、W単位、文節、etc.
- データを公開(共有)する発想の欠如
  - 単に語彙調査のためだけにデータを集めた
  - 報告書執筆後、データは倉庫でホコリをかぶった
  - 1990年頃になっても、著作権処理は無駄という意識があった
  - 国語研の言語データで公開を前提としたのは『太陽コーパス』(1994年開始、公開は2005年)が最初
- 中途半端なコンピュータ利用
  - 1965年に電子計算機を導入(人文系試験研究機関としては初)
  - 調査の規模(延べ語数)は拡大し、複数の語彙表を公開するなどの効果もあったが、集計が早くなり、調査規模が拡大されただけで、理論面では進歩がなかった。NLP的な研究も盛んにおこなわれたが、今日の技術には繋がってはいない。辞書無しのword segmentationなど

コーパス開発

コーパスの要件

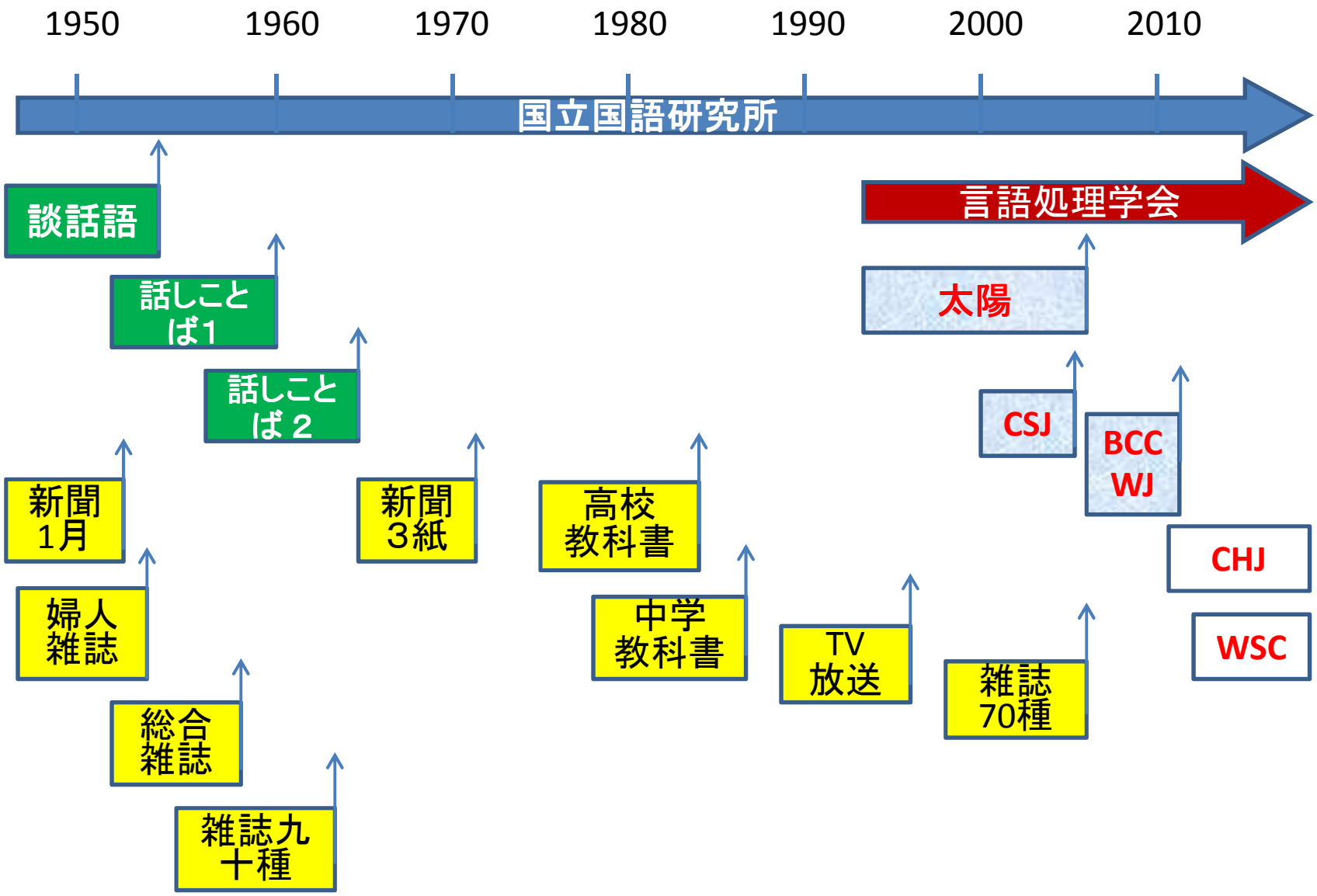
- 代表性：　対象言語変種の全体をとらえている
- 均衡性：　多くの変種をとらえている
- 規模：　　ある程度規模が大きい
- 真正性：　実際に用いられた用例である
- 電子化：　コンピュータで検索できる
- 公開：　　有償無償を問わず誰でも利用できる
- （アノテーション：　検索用情報が付加されている）

国立国語研究所のコーパス

名称（公開年）	対象	規模	特徴
「太陽」コーパス (2005)	総合雑誌「太陽」 1895～1925	推定700万語（短単位）	XML化されたテキストコーパス
『日本語話し言葉コーパス(CSJ)』 (2004)	独話音声中心 (5レジスター)	750万語（662時間）	形態素情報（短単位＋長単位）、節境界、係受け、X-JToBI
『現代日本語書き言葉均衡コーパス (BCCWJ)』(2011)	現代の書き言葉 (11レジスター)	1億500万語	形態素情報（短単位＋長単位）、文書構造、書誌情報
『日本語歴史コーパス（平安時代編）』 (2013)	平安時代文学 (14作品)	73万語	形態素情報（短単位）
超大規模コーパス（2016年公開予定）	Web上の日本語 (1億URL)	300～400億語（予定）	形態素情報（短単位）、文節、係受け

それ以外の言語資源


コーパス
開発センター
Center for Corpus Development, NINJAL
国立国語研究所コーパス開発センターでは、
日本語の全貌を把握するための言語コーパス (language corpus)を構築しています。
国立国語研究所
コーパス
ツール
申込方法
KOTONOHA計画
語彙調査データ
報告書
イベント
ご覧になりたいコーパス名をクリックしてください
日本語歴史コーパス 平安時代編
太陽コーパス
日本語話し言葉コーパス
近代女性雑誌コーパス
現代日本語 書き言葉均衡コーパス
明六雑誌コーパス
超大規模コーパス（開発中）
中古和文UniDic
近代文語UniDic
UniDic
昭和の語彙調査
古典作品
近代語
現代語
コーパス利用申込
Subscription
最新情報
> 最新情報リスト
2014/08/06
書き言葉均衡コーパス
【予告】『現代日本語書き言葉均衡コーパス』DVD版公開一時中止のお知らせ
2014/07/29
お知らせ
「第6回コーパス日本語学ワークショップ」サイトがオープンしました
2014/06/18
お知らせ
コーパス管理ソフトウェア「ChaKi.NET（茶器）」の講習会（於：大阪）を7月31日に開催します
2014/04/17
お知らせ
2014/04/17現在、コーパスの利用申し込みが集中しております。
2014/04/09
お知らせ
第5回コーパス日本語学ワークショップの予稿集を公開しました


1950
1960
1970
1980
1990
2000
2010
国立国語研究所
言語処理学会
談話語
話しことば1
話しことば２
新聞1月
新聞3紙
高校教科書
太陽
CSJ
BCC WJ
CHJ
WSC
婦人雑誌
総合雑誌
雑誌九十種
中学教科書
TV放送
雑誌70種

コーパス開発のこれから

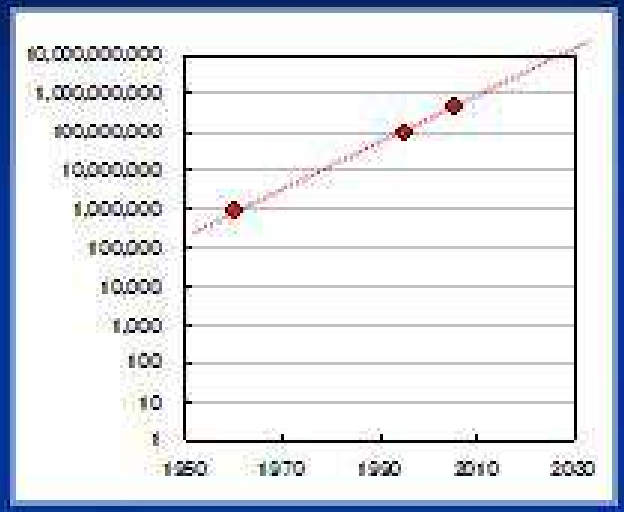
コーパス開発の課題

- 規模の拡大
- レジスターの拡張
- アノテーションの充実
- アノテーション概念の拡張
- コーパス解析手法

規模の拡大

英語コーパスの場合

- 1960 *Brown Corpus*
- 1995 *British National Corpus*
- 2005 *Bank of English*



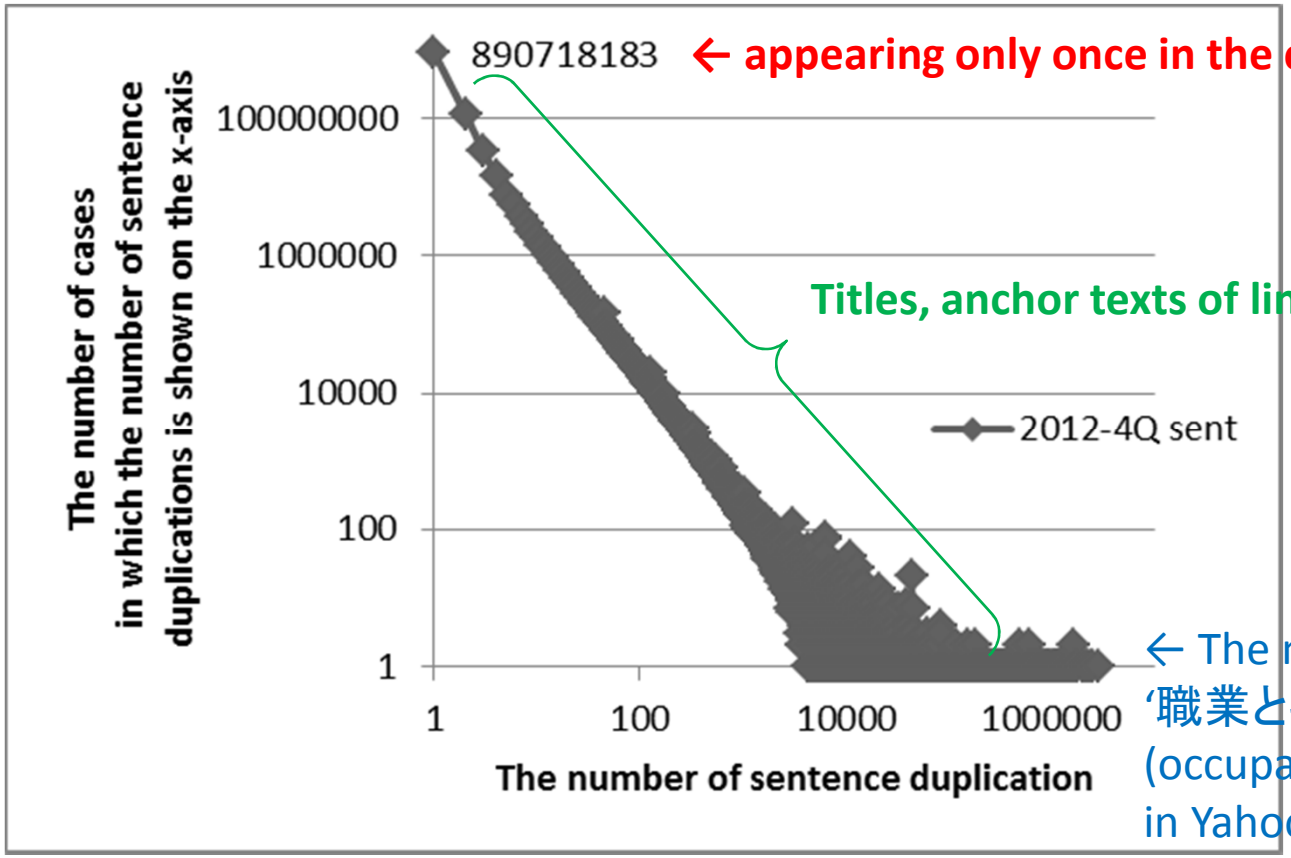
- 過去50年間ほどは、対数軸で線形に増加してきた
- 20年後には100億語を突破
- 日本語もBCCWJの公開後30年で100億に到達するかもしれない

2007年3月特定領域研究「日本語コーパス」公開研究会

超大規模コーパス（構築中）

	2012-4Q	2013-1Q	2013-2Q	2013-3Q
Number of WARC files	814	870	910	905
Number of URLs	61,668,805	58,844,092	61,479,268	57,892,917
Number of Morphemes (w/o sentence extraction)	64,714,650,129	62,077,520,745	63,414,252,638	65,736,027,334
Number of Morphemes (w/ sentence extraction)	33,767,409,441 52.2%	32,651,138,004 52.6%	33,073,991,355 52.2%	30,923,912,566 47.0%
Number of Sentences (Tokens)	2,678,315,774	2,600,122,908	2,659,617,620	2,478,309,312
Number of Sentences (Types)	1,097,011,506	1,048,772,913	1,063,649,324	1,007,771,383

超大規模コーパスにおける文の重複

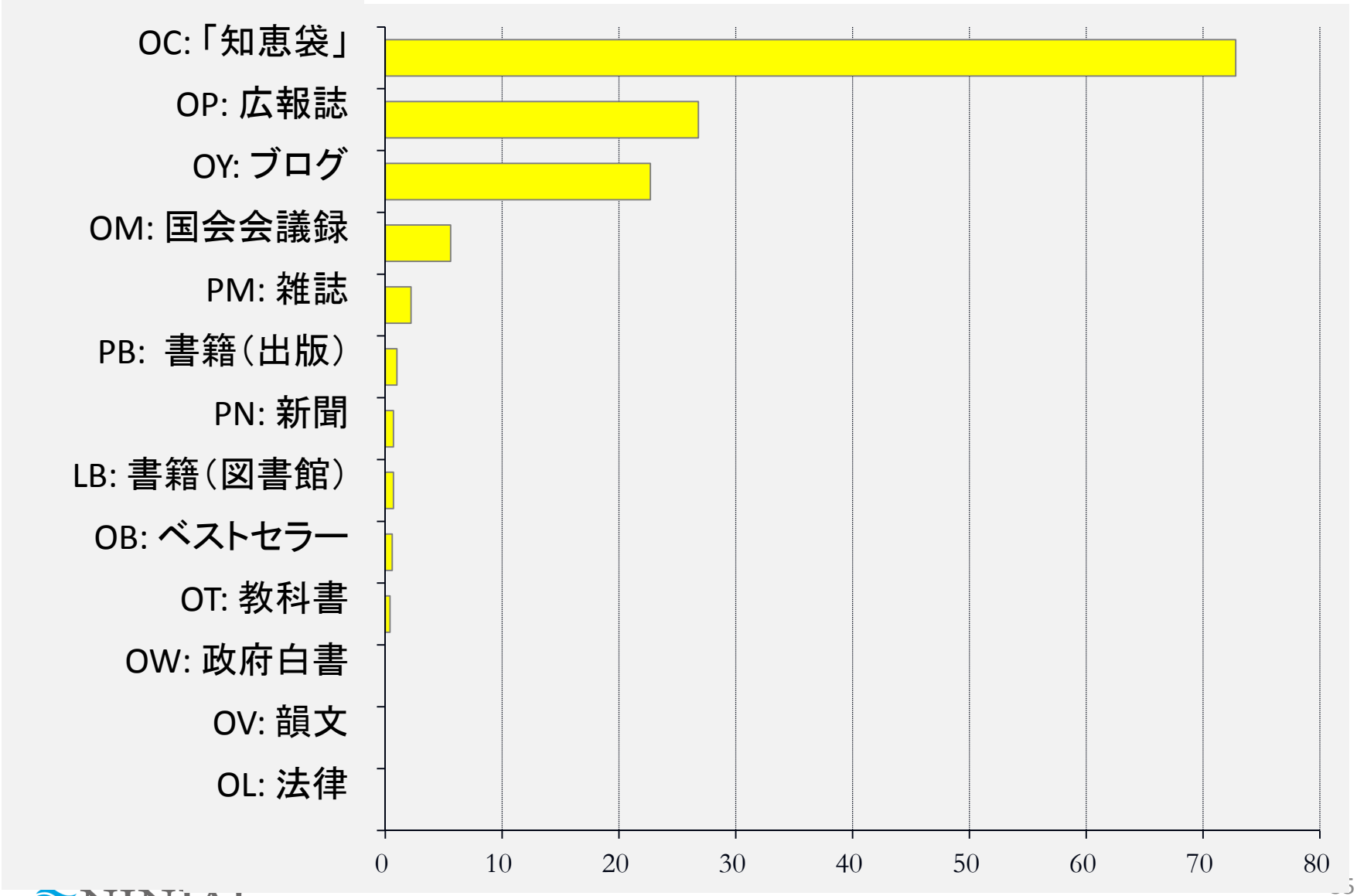


レジスターの拡張

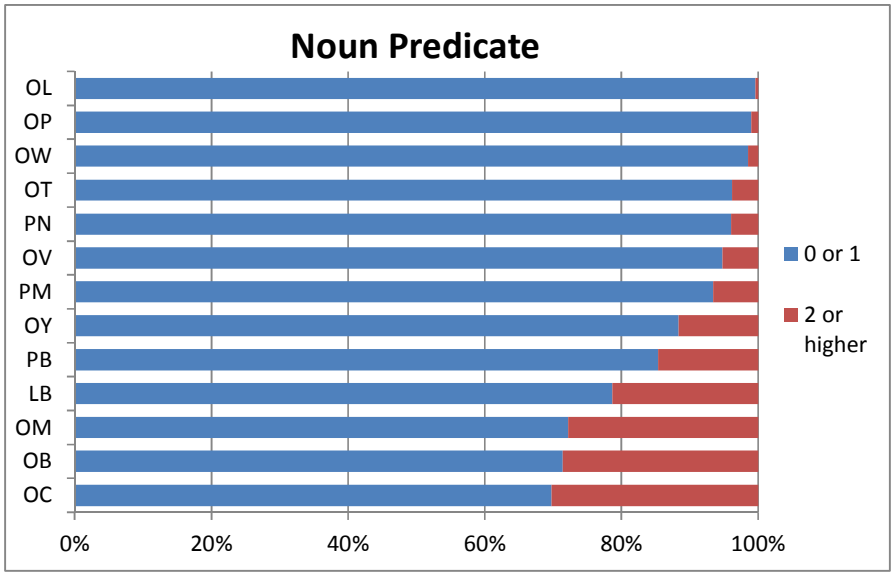
- 超大規模になるとウェブテキストが対象
  - ウェブ全体はひとつのレジスターではない
  - 非常に多くのレジスターの混合物
  - レジスター推定技術が重要

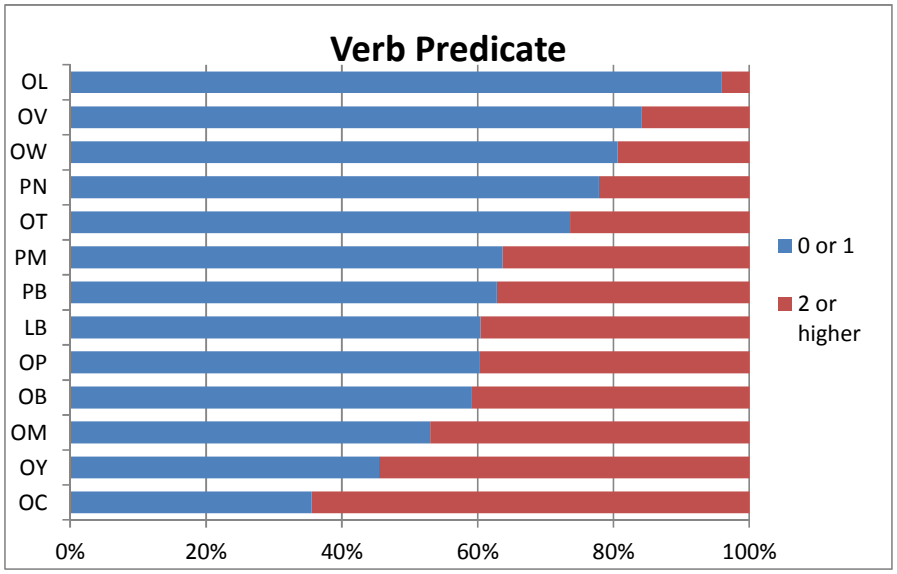
- ウェブではカバーできないレジスター
  - 種々の話し言葉
  - 種々の文芸作品（現代作品）

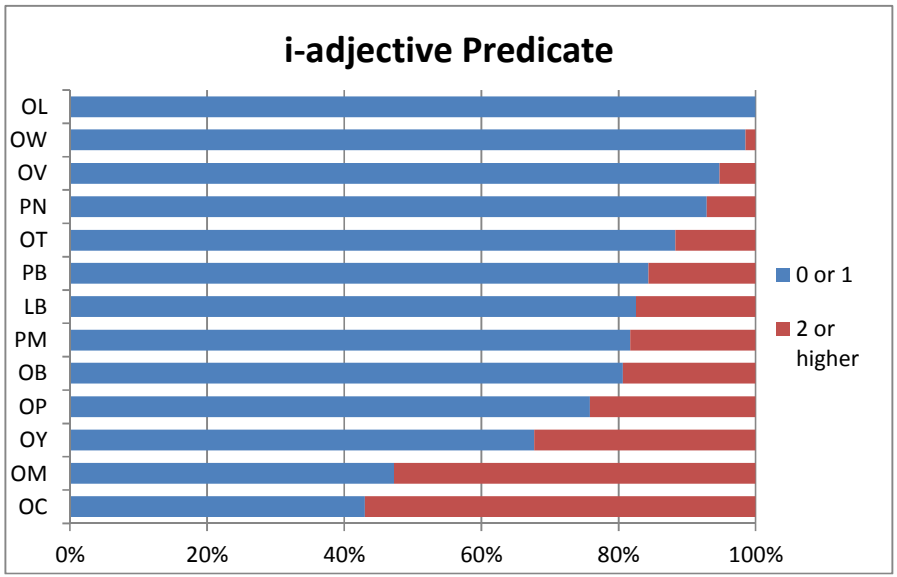
「イ形容詞＋です」述語の生起率（BCCWJ)

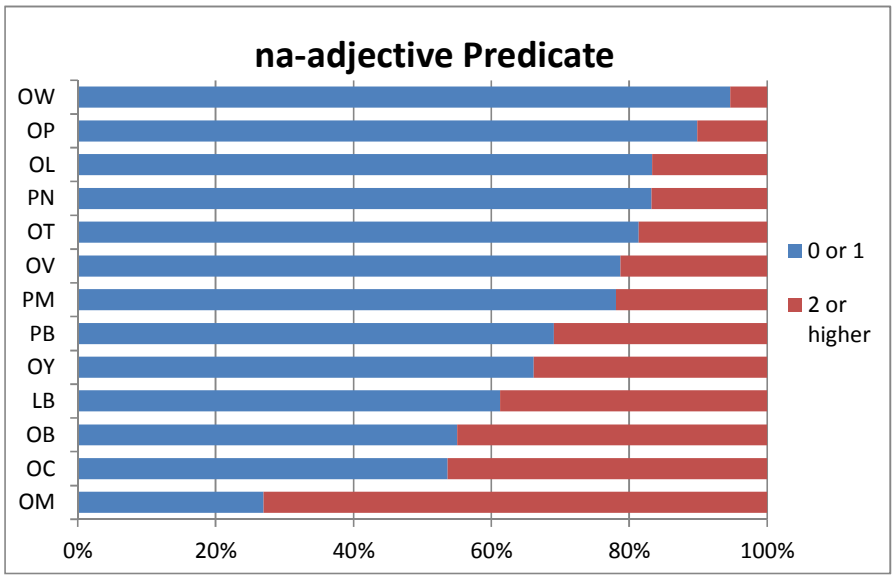


各種述語の複雑さ(長さ)のレジスター差









NINJAL
National Institute for Japanese Language and Linguistics

アノテーションの充実

コーパスの利用価値≈規模×アノテーション

⇒　国立国語研究所共同研究プロジェクト
「コーパスアノテーションの基礎研究」(2010～2015）

作業中のアノテーション

- **文の構造**
  - 文節係り受け構造　【国語研(浅原)、奈良先端大(松本)】
- **文中のセグメント(セグメント系)**
  - 拡張固有表現　【東工大(飯田)】
  - 時間情報表現　【国語研(浅原)】
  - 助動詞「れる・られる」の意味　【国語研(前川・浅原)】
  - 述語境界、節境界　【国語研(前川、丸山】
- **セグメントと文構造の中間**
  - 拡張モダリティ　【東北大(乾)】
  - 否定の焦点　【山梨大(松吉)】
- **述語に関連した文の内部構造(フレーム系)**
  - 述語項構造　【奈良先端大(松本)東工大(飯田)】
  - 日本語フレームネット　【慶応大(小原)】
  - 動詞項構造シソーラス　【岡山大(竹内)】
- **その他**
  - 韻律構造、読み時間情報、等　【国語研(小磯・前川・浅原)】

研究としてのアノテーション

『自然言語処理』21巻2号「コーパスアノテーション―新しい可能性と共有化にむけての試み」

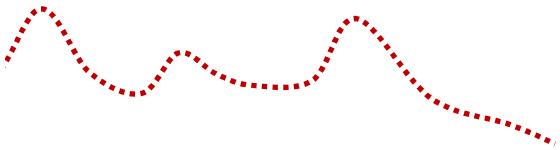
- 投稿14件（後、取り下げ2件）
- 9件採録（採録率75%）

重要だが未着手のアノテーションの例

- 社会言語学的アノテーション
  - 話し手／書き手の属性
    - 年齢
    - 性別
    - 出身地
    - 教育レベル
    - 職業
    - 性格
    - 趣味
    - 人間関係
    - Etc.

アノテーション概念の拡張

- 常識：アノテーションには唯一の正解（真値）がある
  ⇒　カッパ値の高いアノテーションが良いアノテーション
- 常にそうか？
  - X-JToBI（韻律アノテーション）における韻律境界
    - 例：ある部分でピッチレンジがリセットされているかどうか
    - 例：ある箇所で「発話」が終了しているかどうか



  - 局所的にみた場合と大局的に見た場合で解釈が異なる
  - 人間の音声情報処理も同じでは？

⇒ 「分布」としてのアノテーション？

コーパスの解析

コーパスデータの特徴

- 多くの場合に計数データ（ポワソン分布）
- 個人差、レジスター差に意味がある
- 非常に多くの要因が関与（交互作用もあたりまえ）

⇒　頻度主義的な統計解析の限界

- 仮説検定ではなく言語運用のモデル構築が重要
- 階層ベイズモデルなどが魅力的
- ただし言語学者にベイズ統計を教えるのは大変

まとめ（のようなもの）

- 前半では国立国語研究所における言語資源開発の先駆けといえる「話しことば研究」と「語彙調査」の研究を紹介した
- その後、「コーパス」開発の現状を紹介した
- 後半では、これからのコーパス開発の課題を論じた
- 当面（少なくとも10年程度）、国立国語研究所の活動の重点は、言語資源開発におかれると思われる
- 開発と解析の両面で言語処理学会と相携えて前進していきたい

謝辞

本発表資料の一部を提供してくださった、国語研の山崎誠さん、浅原正幸さん、丸山岳彦さんに感謝します

参考文献

Asahara, M., K. Maekawa, M. Imada, S. Kato, and H. Konishi. “Archiving and Analysing Techniques of the Ultra-large-scale Wev-based Corpus Project of NINJAL, Japan”. *Alexandria*, 25 (1) in press.

Maekawa, K., M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka, and Y. Den. “Balanced corpus of contemporary written Japanese”. *Language Resources and Evaluation* 48 (2), pp.345-371, 2014.

Maekawa, K. “Corpus-based phonetics”. In H. Kubozono (ed.) *The Handbook of Japanese Phonetics and Phonology*. Mouton. 2015.

淺原正幸・前川喜久雄「巻頭言：コーパスアノテーション—新しい可能性と共有化にむけての試み—」自然言語処理, 21 (2), pp.95-98, 201

前川喜久雄「コーパス日本語学の可能性—大規模均衡コーパスがもたらすもの—」日本語科学, 22, pp.13-28, 2007.

前川喜久雄「「形容詞＋です」述語の生起要因についての準備的考察」,第1回コーパス日本語学ワークショップ予稿集, pp.211-220,2012.

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.

山崎誠「国立国語研究所の語彙調査の歴史と課題」 http://www.p.u-tokyo.ac.jp/sokutei/pdf/vol06/p168-186.pdf

山崎誠「語彙調査の系譜とコーパス」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.