The Tohoku University 21st Century COE Program in Humanities:Open Lecture Series No.1
東北大学 21 世紀 COE プログラム（言語認知総合科学）第 1 回公開講演会資料

コーパスに基づいた言語研究

日本語話し言葉コーパスの構築に向けて

Heiko Narrog（ハイコ・ナロック）

北海道大学言語文化部

於：東北大学大学院国際文化研究科 2003/1/27

コーパスに基づいた言語研究

日本語話し言葉コーパスの構築に向けて

Heiko Narrog（ハイコ・ナロック） 北海道大学言語文化部

於東北大学大学院国際言語文化研究科 2003/1/27

今日のお話の概要

1　コーパスとは何か
2　コーパスに基づいた言語の研究とはどういうものか
3　話し言葉の特徴
4　コーパスに基づいた研究
5　日本語話し言葉コーパスプロジェクトについて

1　コーパスとは何か

- “a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description” (Kennedy 1998: 1)
- “large and principled collection of natural texts” (Biber 1998: 12)
- アーカイヴ／ELT（電子テキストライブラリー）／コーパス（長瀬1994: 119)
  ELT　「標準的な形式をもったテキストテータベース・・・内容等に関連して一定の取決めはあるが、厳密な選択基準にまで至っていない。」
  コーパス　「 ELTのサブセット。ある目的のために明確な標準に従って大量のテキストデータベースが蓄積された」

コーパスとは何か（続）

- 言語学におけるコーパスの四つの特徴 (McEnery 2002: 29f):
  1）代表性／標本抽出、2）量的限界の有無、3）機械可読性、4）標準的参考資料としての性質
- ある研究分野において重要性を持つ変数について信頼できる頻度などの統計量を得られるように設計されたテクストの集合体

異なる目的のために使用する異なるコーパス

- General Purpose Corpora

■ 翻訳コーパス
■ 同等な構成を持つコーパス(*comparable corpus*); 例：*International Corpus of English*（20か国・地域）
■ Language for Specific Purpose, Special Corpora, 学習者コーパス（Learner Corpora）

コーパスの例

■ BNC (British National Corpus)

- 出版者（Oxford UP、Longman などの出版者と、大学などの研究所の提携 で 1991～ 1994年に構築
- １億語含む、中９０%書き言葉、１０%話し言葉
- 徹底した標本抽出
  ・書誌目録からの無作為標本抽出（テキストの生産側を対象とした標本抽出）やベストセラーリスト、図書館貸し出し人気ランキングなど（読者側を対象とした抽出）を組み合わせるなどする
  ・分野、時代、発行媒体、レベルによる多層偶然無作為抽出
  ・ひとつのテキストは４万語を超えないように切り抜きされる
  ・TEI基準によるマークアップ

BNC（続き）

■ BNCの書き言葉の構成

フィクション（文学的テキスト）25%、非フィクション75%

BNC（続き）

■ BNCの話し言葉部分

- 10,000,000語
- 人口統計的に構成された部分40%、話し言葉の文脈（ジャンル）によって構成された部分60%
- 人口統計的に構成された部分は、124人のボランティアがテープレコーダーを持ち歩いて計1000人ぐらいの話し相手との話しを収録した
- 文脈別の部分は、講議、テレビ・ラジオ放送など、非日常的な話し言葉も含める

Bank of English

■ 出版者（Harper Collins)と大学(Birmingham)が共同で1991年から運営している
■ モニターコーパス　ー　常にテキストが追加される、コーパスの「質」や「代表性」より「量」が求められる
■ 現在（1/2002）４億５千万語の規模（一般利用可能は、そのうち５千６百万語の

み）
- その他：Cambridge International Corpus; CUPとNottingham 大学; ６億語; 完全非公開

COSMAS　コーパス集

- Institut für Deutsche Sprache（ドイツ語研究所）で1993年から運営
- 現在18億4千600万語を収録、中約11億万語が一般利用可能
- 話し言葉は三つのコーパス（計約160万語）が含まれる
- 原則として、オンライン利用のみ可能である

日本語のコーパスは？

（奈良先端科学技術大学院大学（松本裕治））

２　コーパスに基づいた言語の研究とはどういうものか

- 言語研究の二つの流れ
  - 言語能力の研究　　内省によるデータを重視
    例：チョムキー、形式文法論、実際の言語使用には関心を持たない
  - 言語を人間の社会的行動、コミュニケーションの中に位置付ける研究　ー機能言語学、談話・会話分析、（認知言語学）言語の性質が実際の言語使用からしか説明できない

コーパスは、経験的データの調査ならどんな分野でも役立つ

- 語彙論的研究（辞書作成など）
- 文法論的研究
- 社会言語学的研究（方言、言語使用域による差）
- 文体の研究
- 通時的研究
- 言語教育

コーパスに基づいたアプローチ
言語使用の研究の二つの焦点

① 特定な言語構造（語、統語構造、特性）の解明

- 言語構造がどのような文脈に起き、どんな頻度で起きるか
  ⇒　特定なパターンの頻度を調べる
  ⇒　バリエーションを決定付ける文脈を調べる
  - 言語的使用文脈
    ・語彙的使用文脈
    ・文法的使用文脈

- 母語話者でも内省による言葉の認識と知識に限界があり、単に直感に頼れない。大量のデータが必要となる。

言語構造解明の例

little vs. *small*

- *little* はほとんど述語的に使われない（2%以下）；*small*は述語的に使われる（15%以下）。
- *small*は*large*に対応して、主に量を表すのに使われ、*little*は*big*に対応し、具体物・有生物を限定する（以上、Biber 1998）
- *small*は「客観的に」大きさを描写するのに対し、*little*は名詞を同定し、話し手の態度（感情）も表す
  ⇒　Collocations in BNC

②特定な話し手・書き手の言語の特徴の解明

- 特定な著者の言語使用
- 特定な社会層の言語使用
- ジェンダーと言語使用
- 特定なテクストの種類の言語使用
- 例：　新聞、手紙、研究論文の言語　(=register)
  ⇒　個々の言語的特性或いは言語的特性群の使用文脈
  ⇒　言語的特性の共起

例：話し言葉vs.学術の言葉（Biber 1998による）

- 話し言葉の典型的表現：*What you'd have to do, you know, you tell him what you need to know, he'd be able to tell you how to do it*
- ⇒　短縮形、２人称単数、wh-節、「半」助動詞、挿入語
- 学術の言葉の典型的表現：*As has been repeatedly shown, cultural evolution is not a unilinear process, and it is possible that under certain conditions a simpler social formation may emerge out of a more complex one.*

- ⇒　多くの名詞句、限定形容詞、受け身文、外置構造

実際の言語使用における共起のパターンの特徴

- 連続態を成している
- 「常にある」或いは「絶対ない」という観察はまれである。
- 相対的な差がある
  ⇒量的な分析が必要となる

コーパスに基づいた研究の特徴

- 経験的であり、（原則として）帰納的である
- 原則として大型なデータ集合体を利用する
- 分析のために電子計算機を利用する
- 量的な分析を中心とする（ただし、質的な分析との組み合わせは不可欠である。

テクストとコーパスの違い

コーパスに基づいた言語研究の位置

- １）コーパスに基づいた分析は通常、従来の方法論を補うものである。他のアプローチから出た問題を解決するに利用するにすぎない（Biber 1998など）。
  - 構造的分析から生じた問題
  - 理論的考察から生じた問題
  - 直感的観察から生じた問題
- ２）コーパス言語学は使用する方法論が言語の観察に実際に質的な変化をもたらし、言語学への新しい理論的アプローチである（Tognini-Bonelli 2001）。

３　話し言葉vs.書き言葉

話し言葉の優位性の理由

- 書き言葉を持たない言語がたくさんあり、何千年前に世界の言語がほとんど全てそうであった（今でも大半）
- 第一言語習得が話し言葉による
- 文法化は話し言葉による
- ただし、現代社会では書き言葉による影響も侮れない（各種メディア・言語政策

等を通じて）

話し言葉研究のあり方

- 話し言葉の研究は必ず言語使用の研究である
- ただし、（書き言葉と違い）大量のテキストを用意することは困難である
- 表記（書き起こし）の困難さ
- 従来量的な研究より、詳細な、質的研究が多い（会話分析など）

４　コーパスに基づいた研究の成果

- 教育などのための、記述的研究：
  辞書編纂、文法書、教材作成
- 言語の理論的研究

語彙記述、辞書の例

- 名詞dealが最も頻繁に使われる用法は量を示すもの（good, greatとの共起）である。しかし、辞書では、次のようになっている

*deal*の記述の比較（続）

- American Heritage Dictionary (3rd ed) (1992)
- Cambridge American Dictionary (2000)

文法記述の例：LGSWE

理論的研究

他動文の優位性への疑問点Hopper&Thompson 2001

- 446の文からなるアメリカ英語母語話者の会話の他動性の分析
- 結果：文の他動性が非常に低い
- 文の27%しか他動文ではないが、その27%の中もまた、ほとんどが他動性が低い
- 典型的な他動文がほとんど起こらない
- また、Goldberg 1995等で、「言語の文構造の基本をなす」とされる文構造（ditransitive, resultative, caused motionなど）が実際の言語使用では極めてまれである

⇒　文法理論・教育と実際の言語使用のギャップ

他動性のパラメーター

The parameters of scalar transitivity (Hopper and Thompson 1980)

	High	Low
A. Participants	2	1
B. Kinesis	action	non-action
C. Aspect	telic	atelic
D. Punctuality	punctual	non-punctual
E. Volitionality	volitional	non-volitional
F. Affirmation	affirmative	negative
G. Mode	realis	irrealis
H. Agency	A high in potency	A low in potency
I. Affectedness of 0	0 highly affected	0 not affected
J. Individuation of 0	0 highly individuated	0 not individuated

項構造の問題

Thompson&Hopper 2001

- 項構造（argument structure）、格支配などの概念は、ほとんどの統語論で中心的な役割を担う
- しかし、実際の話し言葉で見る限り、項構造には多くの問題がある
  - ー　はっきりした項構造を持たない動詞がある。特に、頻繁に使われれば使われるほどそうなる。
  - ー　項２つあるはずの動詞が頻繁に１つだけの項で起こる（例：forget）⇒２つの動詞？
  - ー　‘P-Words’を伴う動詞；例：*play with, sound like, fit into*
  - ー　V-O compounds；例：*have fun*

項構造の問題（続）

- 項構造が頭の中で予め決まっているものではなく、非常に柔軟であり、使用の中で状況に合わせて変わっていく
- 項構造と頻度が密接な関係にある
- 我々の言語知識は、恐らくきれいに分けられるカテゴリーに分類されたものではない、使用経験の蓄積によるものである

文法関係への疑問

Fujii/Ono 2000, Ono/Thompson/Suzuki 2000

- 話し言葉での対格表示「を」／主格表示「が」の調査（40分（Fujii et al.)／２時間強（Ono et al.）の自然会話の収録から標本抽出）

- 「が」も「を」も、使われうる名詞句のわずか一部にしか使われない
- 「が」は動作主を表示するのにめったに使われない
- 「が」／「を」が使われる時は、それが有標である（無標識が無標）
- ⇒「が」／「を」には、主語・主格／目的語・対格を表示というより談話的な機能（項を有標にする／焦点化する／聞き手の知識で活性化する）がある

抽象論的な認知言語学への疑問

Hallan 2001

- 認知言語学では、path morphemes（pm）の空間的な用法、そして前置詞としての用法が根本的であり、他の用法が派生的であると主張されている（例：Lakoff/Brugman）
- Hallanは、言語習得のコーパス(CHILDESの一部)及びBNCの話し言葉を利用して、*over, on*の用途を分析、説の検証

Hallan 2001（続）

- 大人の話し言葉は元より、言語習得においてさえ、空間的な用法、前置詞としての用法が最初に来るのではない
- 副詞的な、典型的な空間的意味を持たない用法が最初に習得される
- pmごとにその文法を持っている

⇒ 実際の言語使用に根拠を持たない観念論的な認知言語学に問題を提起

構成素構造への疑問（１）

Bybee/Scheibman 1999: *don't*の弱化

- 低い異なり語数＋高い延べ語数が構成素の強い結びつきを生む
- 使用頻度が文法構造を変えて行く
- 頻繁に共起するものはひとつの単位として記憶される。その力が一般的に考えられる構成素構造も変える
- I don't know: [NP[Aux[V]]] > [[NP]AUX]V]: Aux+Vの結びつきよりもNP+AUXの結びつきが強くなっている
- 音声変化と共に、統語構造も変わり、全体の機能も変化している

構成素構造（２）

Bybee 2001: リエゾン

- フランス語のリエゾン（連声）（例：*les[z] autres*）とその分布は構成素構造（constituent structure）では説明できない

- 今まで唯一有効な説明は、２つの要素の共起頻度（*string frequency*）によるものである。更に、変移確率（*transitional probability*）も関与する。（例：*devoir être* vs. *pouvoir être*）

⇒　共起頻度が心的表示に直接影響をするし、頻繁に共起する語は、構成素の境界線を超えても一つの固まりとして記憶される

⇒ 更に、英語の*I'll, I'm*の例

問題提起

- How [are] linguists led to impute structure to any sequence of forms if not on the grounds of their prominence in usage and memory, that is, their usefulness in discourse reflected in their frequency. In other words, what are the alternatives to frequency as an explanation for structure and regularity in language? (Bybee 2001a: 19)
- 頻度以外には、言語の構造と規則性を説明するパラメータがありうるのだろうか。

方法論的問題点

- 話し言葉の優位性の信念は明らかだが、その裏付けと、話し言葉と心内語、書き言葉の関連についてあまり深く考察されていない
- 今紹介した多くの研究は、ごく少量の話し言葉に基づいており、その中でも「代表性」を保証するために何らかのバランスをとろうとしていない

４日本語話し言葉プロジェクト

- 従来の話し言葉研究（特に日本語）はほとんど質的
- 本プロジェクトでは量的な研究を可能としながら、質的な研究もできるようにしたい

⇒ 文法研究、語彙研究、談話研究、社会言語学的研究の各分野で役立つようにしたい

他の話し言葉プロジェクト

- 国立国語研究所の『日本語話し言葉コーパス』プロジェクト（2004年春公開予定）ー　学会発表（モノローグの部分のみ）・模擬講演会を収録
- 名古屋大学のコーパスプロジェクト日本語話し言葉コーパス。2004年公開予定（科研費）
- 小規模なコーパス：ACI Hayama による「インタビュー形式による日本語会話コーパス」（2002～2006年、科研費／北九州市立大学と同一？）（母語話者と非母語話者の比較／音声中心）、『主婦の一週間の談話資料』、『女性（男性）のことば・職場編』等々)

本プロジェクトの特徴

1. 自然会話を対象とする
2. 大規模（100時間）を目的としながらも高い質の表記をする
   ⇒談話・会話分析的研究も可能とする
   ⇒DuBois (1993) に基づかせる
   表記マニュアル
3. TEI形式化を考えている

規模

- 100時間以上（100時間は英語なら大体80万語になる；1000時間以上の表記時間が想定される）

⇒ LLC（1959~1974）は、50万語

⇒ BNCの話し言葉は、1000万語

量と質を合わせた話し言葉の研究

- He&Kennedy 1999: LLCに基づいてTurn-Bidding（発言権交代への働きかけ）の研究
- その言語的手段や頻度と、社会言語学的な変数と関連を詳細に分析

公開／利用

- 名古屋コーパスプロジェクト
- MICASE
- COSMAS

分析ツール

- 英語など欧文に関してConcordancerなどが豊富にあり、語彙的共起関係と頻度情報を調べるのに役立つ
- 日本語は簡単なツールが色々
- 量的分析には分かち書きが決定的

作者連絡先　narrog@ilcs.hokudai.ac.jp