コーパスという生態系に語句の意味を探る

海洋生物学者が海に生物の生態を探るように

黒田　航
情報通信研究機構 けいはんな情報通信融合研究センター
09/19/2005

JCLA 5 (2005) Workshop
「コーパス利用とこれからの認知言語学」

発表のあらまし

- 用法基盤の姿勢 (Usage-based Approach) の考え方について
  - 観察的妥当性の重要性
  - コーディングの必要性
  - コーパス言語学を越える必要性
- コーパスの賢い利用法
- 語の用法の認定のための手法
  - “コーパスは語の生態系である”というアナロジーの精緻化

U sage-ba sed m o del of la ngu age
Where do int ui tions come from?
H ow to de sc ribe la nguage usage?
Aga inst the “P lato nist” vie w of lan guage
emp irici sm and
data
obs erva tio n

用法基盤の考え方

出発点

- 用法基盤主義は現在のところカケ声倒れ
  - 認知言語学では今までのところコーパス事例の綿密な分析に基づいた説明は見られない
- 例外
  - Barlow, M. and Kemmer, S, eds. 2000. Usage-based Models of Language. CSLI.
  - Stefanowitsch, A. & Gries, S. Th. 2003. “Collostructions.” International J. of Corpus Linguistics 8 (2), 209-243.
  - 早瀬尚子. 2004. “コーパスと認知言語学: 英語における姿勢維持動詞の意味と構文の発展について”. 英語コーパス学会でのシンポジウム
- でも，なぜ??

観察的妥当性の根本的重要性

- あまりに多くの言語研究で観察的妥当性が満足されていない
  - 説明的妥当性より前に記述的妥当性，記述的妥当性の前に観察的妥当性
  - 理論は観察を良い方にも悪い方にも導く
    - 観察の理論負荷性 (Hanson 1958)
- 問題
  - 観察的妥当性は，どのようにして満足されるか?

そもそも，観察とは何か?

- 観察とは何をどうすることか
  - 言語学では観察が何であるかの説明がない
  - 観察とは観測データが “値” となるような “特徴” 集合の明示
- 特徴の例
  - 生物個体の {体長, 体重, 体色, 行動パターン, ...}
    - 体長-体重, 体色と雌雄の相関
    - 行動パターンと性別(雌雄)の相関

観察には“技”と“根気”が必要

- 観察にはしばしば
  - 特殊な観察支援装置
    - 顕微鏡や望遠鏡
  - 特殊な観察支援技術
    - 染色法 (特徴マーキングの一種)
    - タグ=認識札づけや発信器による個体 (ID) マーキング
      - 識別困難な状態にある個体の認識
- が必要で，なおかつ訓練も必要
- 観察は手間がかかり，根気が必要なるモノ

何で手間をかけるのか?

- 生データはそのままでは何も語らない
  - データは解釈されて初めて意味をもつ
  - ただし解釈は可能な限り研究者の恣意を排除した形でなされる必要がある
  - 妥当な解釈を保証するためには妥当な加工が必要
- コーディングはデータに内在するが見えにくい特徴の可視化を支援
  - 恣意性を避けながら重要な特徴を浮き彫りにする加工法が特徴コーディング

確かに大変な作業だが ...

- 用法基盤主義がカケ声倒れに終っている最大の理由は，手間のかかり根気の必要な観察を
  - 「大変だから止めとこう」と言う人が
  - 「大変だからやってやろう」と言う人より
- 数が多いことにあるのでは?
- でも，いったい
  - 妥当な観察なしに，どうして妥当な記述が成立??
  - 妥当な記述なしに，どうして妥当な説明が成立???

用法基盤主義が広まらないワケ

- Langacker 流の用法基盤モデルの難点
  - 用法という概念に明示的な操作的定義が必要
    - そもそも用法 (usages) の定義がない
    - Usage Event って図に書けばいいってもんじゃない
- “理論” 言語学者の多くは十分な観察が前提とならない “直観基盤” の分析が好き
  - 十分な量の観察をしていない人に Generalization Commitment を云々する資格はない
  - 理論家ぶって自分の手抜きをメタ理論的に正当化するのは止めよう

観察対象の定義の二つの流儀

- 直観基盤 (intuition-based) の内観的 (introspective) 記述モデル=アプローチ
  - 作例が記述の対象; 実際の使用例は参考値
- 観察基盤 (observation-based) の外観的 (extrospective) 記述モデル=アプローチ
  - 実際の使用例が記述の対象; 作例は参考値
- 重要な点
  - これらのアプローチは本来は相補的なもので目的によって使いわける必要がある

用法基盤の姿勢とは?

- 意味記述を
  - (コーパス言語学がそうするように) 生のデータ/分布に帰着させるのではなく
  - (Goldberg, Lakoff, Langacker らがそうするように) 上位にある，実質の足りない，抽象的な(イメージ)スキーマに帰着させるのでもなく
- なるべく具体的で個別的で(しばしば雑多な)下位スキーマ群に帰着させようとする姿勢
  - “記述のための記述” と “説明のための説明” の両者を排除する

コーパスを研究に利用する意味

- 事例を網羅的，体系的に集めることで，観察レベルでのバイアスが回避可能
  - 作例中心の研究は観察のレベルでのバイアスを避けがたい
  - これは生成言語学で顕著だが，認知言語学でも是正されているわけではない
- ただしコーパスの利用にはそれ自体に特有の制約もあるのは知っておくべき

コーパス言語学を越えて

- コーパス言語学の最大の成果は優れた辞書群
- これが意味すること
  - 認知言語学がコーパス言語学を「越える」のに必要なのは,
    - 「彼らのやっていることは単なる分類だ」と相手を見下すことではなく
    - 辞書編纂者に可能な以上の精度で語，ならびに語より大きな単位の意味の記述を体系的に行なうこと
    - いわゆる「説明」はこの後
- これは直観のみでは実現不可能

Co rp us Lin guis tics a nd B ey ond
Fr om Na ïv e, Int ui ti on-bas ed C ogn
it ive Li ngu istics to
Data-driven, Truly U sag e-b ase d Co
gn it ive Lin guis ti cs

コーパスの賢い利用法

いわゆる「統計」処理の限界

- BNC コーパスを KWIC 検索して何千〜何万という実例が得られた．さてどうする?
  - 漠然と実例を眺めているだけで何かがわかるワケではないので，何かしなければならない
  - コーパス言語学の常套手段は「とにかく数える」
  - 結果は意味的に未解読の共起関係の一覧
- 問題
  - 頻度を調べるのはいいが，問題は何の頻度を，何のために数えるのか

単なる統計を越えて

- コーパス言語学 (と計算言語学) は典型的に
  - 統計処理に奔り
  - 直観に基づく意味分析を排除，あるいは軽視する
- 重要なのは直観を排除することではなく，それを十分な量の観察と融合させること
- 統計は有用だが万能ではない
  - コーパスには正例しかないので，正例と負例の境界を正しく記述できるとは限らない
  - 主観性を排除するという理由から，直観に基づく意味分析を排除するのは方法論的に誤り

W hy c ode d ata an d how?
Why us age-ba sed ap pro ach?
W hy Chom skian li nguists are b li nd to dat a?
Is ling uis tics an em piri cal scie nce, an d if so, how is it?

コーパスに意味の生態を探る

意味の “生態” を探る際の困難

- 語句の意味は暗黙的で(なかなか)目に見えない
  - 語の意味は非言語的情報として与えられる
- 意味がイメージだというのは認知言語学の主張の一つだが，その妥当性は鵜呑みされるべきでなく，実証的に示されるべき
  - コトバの意味の(極く)一部はイメージ(スキーマ)かも知れないが，全部がそうだというワケではないし，
  - イメージがコトバの意味の本質的に重要な部分かどうかもわかっていない

意味をどう “特定” するか

- これこそが意味記述の根本問題
  - 意味がどう認定されているかは，言語学の教科書では自明と見なされ，真剣に議論されることが少ない
  - 「自明」だと思われていることは，しばしばもっとも説明困難
- 私が生態学の研究手法に基づいて提案する体系的意味記述の方法
  - コーパスから事例を網羅的に収集し，それを手間を厭わずコーディングし，その結果の(多変量)解析を通じて帰納的に意味のタイプを発見する

語の意味と用法との関係

- 作業仮説
  - 語の意味はトークンとしての語にではなく，タイプとしての語の用法に現われる
  - 語の用法は生起環境と相関によって特定できる
  - 用法をうまく特定できれば，それに基づいて意味のタイプを特定できる
- これで基本はよいとして，この案をどう実装する?

用法への生態学的アプローチ

- 有益なアナロジー
  - コーパスを生態系 (例えば海) に
  - 用法 = 意味を生物種に見立て
  - 用法 = 意味を発見, 観察, 記述する方法を生態学から援用する
- 注意
  - これは確かにアナロジーだが単なるアナロジーではない
  - コーパス言語学がやっていることの実質はこれ

「襲う」のコーディング例

	L_CONTEXT	KEY	KEY_F	R_CONTEXT	SUBJECT	SUBJECT_	OBJECT	OBJECT_	F_L1	F_L2	F_L3	COMMENT
7	地雷は無差別に人を	襲う	active	兵器であり、児童が犠牲者となるケースがきわめて多い。	地雷	兵器[+military]	人	人	人為:兵器による攻撃[+military,+metaphoric]	打撃[-human,+military,-animate,-intentional,+concrete]		[-human]-->[+human]: personification
8	それから約四十年後の五七年、再びインフルエンザが世界を	襲った	active	。	インフルエンザ	疫病	世界	地域	自然:疫病の流行	打撃[-human,-animate,-intentional,-concrete]?	自然災害の発生	打撃系
9	中世ヨーロッパを	襲った	active	ペストのような伝染病が大流行している訳ではない。	ペスト	疫病	中世ヨーロッパ	地域[+time]	自然:疫病の流行	打撃[-human,-animate,-intentional,-concrete]?	自然災害の発生	打撃系
10	合法的な就労に道が開かれている日系外国人にも景気後退の波が容赦なく	襲い	active, compound	かかろうとしている。	景気後退の波	打撃	日系外国人	人[+group]	自然:活動への打撃	打撃[-human,-animate,-intentional,-concrete]	異変の発生	打撃の発生 AS 津波の発生
11	東独を含め、東欧諸国に民主化の波が激しく	襲い	active, compound	かかった。	民主化の波	津波[+metaphoric]?	東欧諸国	場所[国]	自然:活動への打撃	打撃[-human,-animate,-intentional,-concrete]		異変 AS 津波; 国 AS 活動体 AS 生体?
12	この動きは、子弟を欧米に留学させている一般大衆をもドル買いに走らせ、バーツ売り・ドル買いは大波となってタイ市場を	襲い	active	、タイを国家破産の状態に追い込むかの勢いとなった。	バーツ売り・ドル買い[+大波となって]	活動[+simile]	タイ市場	活動体?	自然:活動への打撃	打撃[-human,-animate,-intentional,-concrete]		打撃系
13	また、市場は先週のドイツ連銀の公定歩合引き下げ見送りを見て、通貨統合を進めるうえでカギとなる独仏協調にひびが入っていると判断、欧州通貨に	襲い	active, compound	かかった。	市場?	活動場所[+metonymic]?	欧州通貨	価値体系	自然:活動への打撃	打撃[-human,-animate,-intentional,-concrete]		打撃系
	不良債権の処理や金融再編など、金融システムの安定を図る過程では、信用収縮	襲い	active	、過渡的な摩擦を起こすこ	信用収縮によるひ	打撃	経済の弱	経済活動	自然:活動	打撃[-human,-animate,-		

生態系のアナロジーの帰結 1/2

- 共起制限とは要するに，語の用法という “種” の生態学的特徴
  - 種に相当するのは語ではなく，語の用法であり
  - これが “語義” (sense) に対応する
  - 種 s = 語義 sense の個体が一つ一つの使用 (usage event)
- これが意味すること
  - 語や lemma 概念の記述力の過剰/説明力の不足
    - 語は意味上は種ではなく “属” や “科” のような超種タクソン
  - Lexical Units (FrameNet after Melcuk) の妥当性

生態系のアナロジーの帰結 2/2

- コーパス事例の検討 (外観) = 野外での観察
  - 対象の自然状態での自然な挙動が観察可能
  - 挙動の説明要因は統制困難
- 作例の検討 (内観) = 実験室での観察
  - 挙動の説明要因の統制が可能
  - 観察は自然な挙動を反映したものとは限らない
- 内観か外観かの二者択一ではなくて，両者の上手な組みあわせが必要

W hy c ode d ata an d how?
Why us age-ba sed ap pro ach?
W hy Chom skian li nguists are b li nd to dat a?
Is ling uis tics an em piri cal scie nce, an d if so, how is it?

まとめ

コーパスの利用価値

- 注釈なしの生コーパスはそれほど有用ではない
  - 十分な量のデータを収集することは必要だが，集めたデータを漠然と眺めるだけでは十分ではない
  - 特にコーパスに意味の生態を探りたい場合，語がトークンとなる意味のタイプが指定されていない限り，頻度分析はそれほど有効ではない
- データ自体に何かを「語らせる」ためにはコーディングが不可避
  - 研究に重要な情報のみをうまく取り出す必要がある

コーパスに意味の生態を探るなら

- 意味タイプを同定するための作業が必要
  - 金丸の発表で省力化の手法の紹介
- そのための技法の一つが特徴コーディングで，その結果に基づいて興味深い記述的一般が得られる
  - 黒宮の発表で実例の紹介
- 多変量解析による既成の説明の検証も可能
  - 李・濱野の発表で実例の紹介

謝辞

以下の方々との議論が参考になりました

飯田 龍 (奈良先端大学)

竹内 和広 (情報通信研究機構)

中本 敬子 (京都大学)

京大山梨研究室の自主ゼミ「統計入門」参加者

以下の方から技術的な支援を頂きました

内山 将夫 (情報通信研究機構)