コーパス調査入門
イントロダクション

4/15 (水)

淺尾仁彦

自己紹介

- 淺尾仁彦 (あさおよしひこ)
- 言語学、とくに、形態論・意味論
- 京都大学 > ニューヨーク州立大学バッファロー校
- 名古屋には来たばかりです。今週が（ほぼ）初めての日本での授業です

コーパスとは

- 厳密な定義は特に必要ないが
- **実際に使用された言語のデータ**を、**電子的**に、**研究（／教育／技術開発）目的**で収集したもの
  - 典型的にはテキストデータ
  - しかし、音声や動画のコーパスも存在する（ジェスチャー、視線、手話…）

言語研究のデータ

- **内省**：言語に対する話し手の判断
  - 研究者自身の内省
  - インタビュー、アンケート
- **コーパス**：実際に使用された言語
  - 書かれたものや、録音・録画など（歴史的な資料を含む）
  - このうち、電子的に利用可能で規模の大きいものを典型的に**コーパス**と呼ぶ
- その他の手法（心理実験/神経科学実験など）

コーパスを使うと何ができるか

コーパスの得意分野

- 直観では判断しにくいような微妙な意味・用法の違いを発見する
- また、その個人差、ジャンルによる差、歴史変化などを調べる

「風景」と「光景」

- 留学生に「「風景」と「光景」は使い方が違いますか？」と聞かれました。なんと答えますか？
- http://nlb.ninjal.ac.jp/

さ入れ言葉

- ある人が「以上で終わらさせていただきます」と言っていた。「終わる」の使役は「終わらせて」では？
- このような言い方は広まっているのか？いつ広まったのか？人による差はあるのか？

さ入れ言葉 (佐野 2008a, b)

- 日本語話し言葉コーパス (CSJ) では、さ入れ言葉は 2.73%
  - 「-ていただく」が続くときは 13.64%
  - 男性の「さ入れ率」3.68%、女性は 1.30%
- 国会会議録では、1980年代までは数例しかない。1990年代に95例、2000年代に203例と激増

少しだけ歴史

Zipf

- Zipf (1935)

チョムスキー

- 20世紀前半：科学性・客観性の観点から、コーパスから機械的に文法を記述する方法が追究される（構造主義言語学）
- Chomsky (1957) *Syntactic Structures*: 「コーパスによる頻度では言語は記述できない」
  - “Colorless green ideas sleep furiously”：文法的だがコーパスに自然に現れるとは考えにくい
  - Competence vs. Performance (Chomsky 1965)

コーパス (≒実例) と内省 (≒作例) の違い

コーパス	内省
• 正例しかない（ただし、誤植・書き間違いを含むかもしれない） • 大勢の話者からデータを集めたり、統計的な差を見るのが比較的容易 • 遠い過去のデータもある • 自然な言語使用の観察	• 正例（容認可能なもの）と負例（容認できないもの）の両方を収集できる • 大勢からデータを集めるのは手間 • 遠い過去のデータは少なく、生きている（言語学者の質問が理解できる）話者がいないと調査は不可能 • 不自然な状況でのメタ判断

コーパス (≒実例) と内省 (≒作例) の違い

- 実例と作例、両者は「どちらが優れている」というものではなく、別々の種類のデータであり、どちらも解明の対象である
  - コーパスにしても内省にしてもいろいろな要因によって左右されている現象であり、どちらかが「本質」である、などと考えるのは誤り
- 研究目的に応じて、コーパスを使うのがよいか、内省を使うのがよいかが決まる

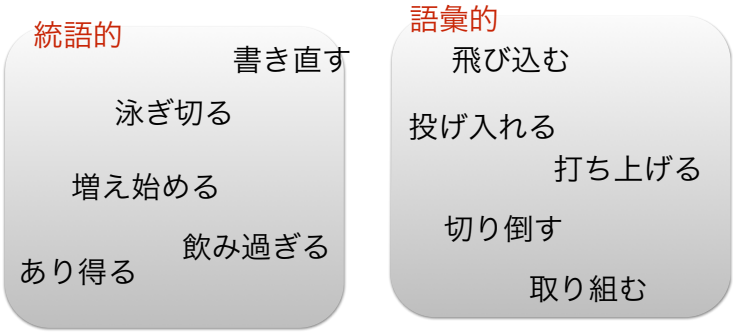
（私の）研究例

日本語の複合動詞

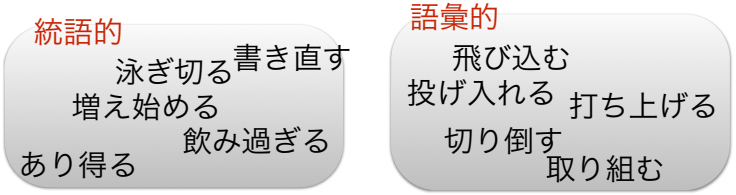
- 私自身は文法と頻度の関連に興味をもってきました
- コーパスで単に「こういう表現の頻度が多い・少ない」ということを調べるのではなく、文法現象とコーパスにおける頻度がどのように関連するかということ自体に興味をもってきました

日本語の複合動詞

- 2種類の複合動詞（統語的複合動詞、語彙的複合動詞）(影山 1993)



文法的違い



そうし始める、そうし過ぎる

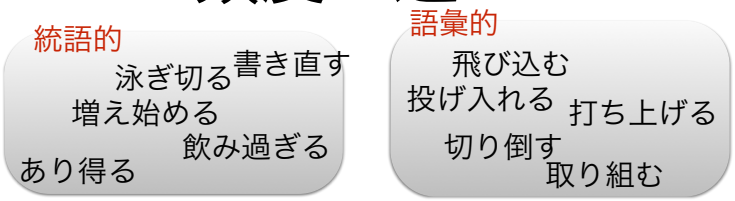
存在し得る、清書し直す

*そうし倒す、*そうし直る

*ジャンプし込む、

*発射し上げる

頻度の違い



- 新聞コーパスから計算した生産性

始める .069	上げる .008
得る .040	入れる .004
直す .018	込む .003
切る .012	組む .000

記憶のされ方の違い？

- 「食べ始める」は「食べる」＋「始める」に分解して理解・産出している
  - したがって「ググり始める」など生産的に作ってもよいし
  - 「私が食べ始めたのを見て、Aさんもそうし始めた」のように、代用形を使うこともできる

記憶のされ方の違い？

- 「切り倒す」は（「切る」＋「倒す」で理解できそうだが、それでも）「切り倒す」のまま暗記している
  - したがって「*切断し倒す」などは（意味が理解できるにもかかわらず）作れないし
  - 「*私が切り倒したのを見て、Aさんもそうし倒した。」もダメ。

コーパスの使われる様々な分野

社会学

- 2001年と2004年のインターネット使用状況と学歴について、「インターネットから何を連想しますか」と聞いた結果の分析 (樋口)

計量文献学

- 統計から、ベーコンとシェイクスピアの同一人物説を否定

自然言語処理 (NLP)

- たとえば機械翻訳
- かつては手作業で辞書と文法を用意していた
- 現在では対訳用例を大量に蓄積し、確率的にもっともありそうなものを答えとして出す手法が主流

授業の進め方

出席・参加

- 毎回授業の終わりに（簡単な）リアクションペーパーの提出を求めます。内容に関してはその都度指定します。
- 議論に積極的に参加してください。
- 学期のなかほどで、論文の紹介をお願いします。

宿題

- 授業の内容を踏まえ、パソコンを操作して具体的にデータを検索するなどの課題に取り組みます。(おそらく 5, 6, 7 月に1回ずつ程度)
- 具体的な内容については追って公開します。

期末レポート

- 期末レポートとして、実際にコーパスを用いて分析をします
- レポートのトピックは事前に相談が必要です（授業のなかで議論の時間を取るかもしれません。詳細は追って連絡します）
- 最終回近くに、レポートの内容について発表します。

リアクションペーパー

- 紙に、名前・学年・専攻・学籍番号を書いてください。
- この授業に何を求めるか、一言コメントを書いてください。