認知言語学とコーパス分析

李在鎬

背　景

● **理論言語学の盲点**
- 作例による一般化は本質的に循環論⇒実証性の乏しい考察⇒応用性の欠如
- 認知言語学も例外ではなく、問題視する声が出始めている現状

● **コーパス分析が切り開く可能性**
- 自らの分析モデルを評価する良い手段⇒意味研究への貢献
- 言語単位の認定に関する新たな方法論⇒構文文法、語彙意味論への貢献

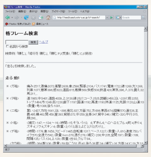
！どんなデータベースが使えるのか
!! 何をどうやって用例を集めるのか
!!! 用例を集めたところでどう分析するのか

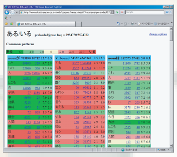
どんなデータベースが使えるのか

● **ウェブコーパス(≠ネット検索)の位置づけ**
- ウェブ上のデータをコーパス化したもの
- 自然言語処理の分野で利用

→格フレーム辞書、Sketch Engine、ウィキペディアなど
- 長所：大規模でかつ、著作権フリー、簡単、無料
- 短所：書き手が不明,メンテなし,ノイズ多し

→長所を生かし、コロケーション調査などでは有効利用可能

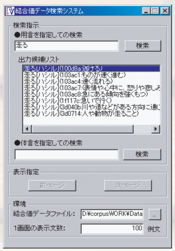




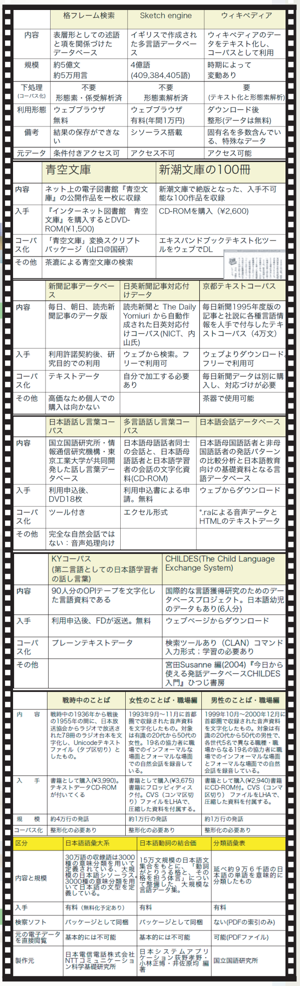
● **書きことばコーパスの現状と位置づけ**
- 新聞、小説など複数のジャンルのテキストあり
- 代表性の問題をクリアするのが難しい
- 利用には制限が多いため、要確認

→日英新聞記事対応付けデータ、日本語均衡コーパス（国研）、青空文庫
- 長所：書き手・出典が明確、表現の規範性が高い
- 短所：手間・コスト高、データのバイアス強

→複数のデータを比較評価しながら利用するのが望ましい



＊リンクは配布資料参照



	格フレーム検索	Sketch engine	ウィキペディア
内容	表層形としての述語と項を関係づけたデータベース	イギリスで作成された多言語データベース	ウィキペディアのデータをテキスト化し、コーパスとして利用
規模	約5億文 約5万用言	4億語 (409,384,405語)	時期によって 変動あり
下処理（コーパス化）	不要 形態素・係受解析済	不要 形態素解析済	要 (テキスト化と形態素解析)
利用形態	ウェブブラウザ 無料	ウェブブラウザ 有料(年間1万円)	ダウンロード後 整形(データは無料)
備考	結果の保存ができない	シソーラス搭載	固有名を多数含んでいる、特殊なデータ
元データ	条件付きアクセス可	アクセス不可	アクセス可能

	青空文庫	新潮文庫の100冊
内容	ネット上の電子図書館『青空文庫』の公開作品を一枚に収録	新潮文庫で絶版となった、入手不可能な100作品を収録
入手	『インターネット図書館　青空文庫』を購入するとDVD-ROM(¥1,500)	CD-ROMを購入（¥2,600)
コーパス化	「青空文庫」変換スクリプトパッケージ（山口@国研）	エキスパンドブックテキスト化ツールをウェブでDL
その他	茶漉による青空文庫の検索	

	新聞記事データベース	日英新聞記事対応付けデータ	京都テキストコーパス
内容	毎日、朝日、読売新聞記事のデータ版	読売新聞と The Daily Yomiuri から自動作成された日英対応付けコーパス(NICT、内山氏)	毎日新聞1995年度版の記事と社説に各種言語情報を人手で付与したテキストコーパス（4万文）
入手	利用許諾契約後、研究目的での利用	ウェブから検索。フリーで利用可	ウェブよりダウンロード。フリーで利用可
コーパス化	テキストデータ	自分で加工する必要あり	毎日新聞データは別に購入し、対応づけが必要
その他	高価なため個人での購入は向かない		茶器で使用可能

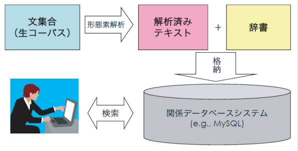
	日本語話し言葉コーパス	多言語話し言葉コーパス	日本語会話データベース
内容	国立国語研究所・情報通信研究機構・東京工業大学が共同開発した話し言葉データベース	日本語母語話者同士の会話と、日本語母語話者と日本語学習者の会話の文字化資料(CD-ROM)	日本語母国語話者と非母国語話者の発話パターンの比較分析と日本語教育向けの基礎資料となる言語データベース
入手	利用申込後、DVD18枚	利用申込書による申請。無料	ウェブからダウンロード
コーパス化	ツール付き	エクセル形式	*.raによる音声データとHTMLのテキストデータ
その他	完全な自然会話ではない：音声処理向け		

	KYコーパス (第二言語としての日本語学習者の話し言葉)	CHILDES(The Child Language Exchange System)
内容	90人分のOPIテープを文字化した言語資料である	国際的な言語獲得研究のためのデータベースプロジェクト。日本語幼児のデータもあり(6人分)
入手	利用申込後、FDが返送。無料	ウェブページからダウンロード
コーパス化	プレーンテキストデータ	検索ツールあり（CLAN）コマンド入力形式；学習の必要あり
その他		宮田Susanne 編(2004)『今日から使える発話データベースCHILDES入門』ひつじ書房

	戦時中のことば	女性のことば・職場編	男性のことば・職場編
内容	戦時中の1936年から戦後の1955年の間に、日本放送協会からラジオで放送された78冊のラジオ台本を文字化し、Unicodeテキストファイル（タブ区切り）としたもの。	1993年9月～11月に首都圏で収録された音声資料を文字化したもの。対象は有識の20代から50代の女性。19名の協力者に職場でのインフォーマルな場面とフォーマルな場面での自然会話を録音している。	1999年10月～2000年12月に首都圏で収録された音声資料を文字化したもの。対象は有識の20代から50代の男性で、各世代5名で異なる職種・職場からなる19名の協力者に職場でのインフォーマルな場面とフォーマルな場面での自然会話を録音している。
入手	書籍として購入(¥3,990)。テキストデータCD-ROMが付いてくる	書籍として購入(¥3,675)書籍にフロッピィディスク付。CVS（コンマ区切り）ファイルをLHAで、圧縮した資料を付属する。	書籍として購入(¥2,940)書籍にCD-ROM付。CVS（コンマ区切り） ファイルをLHAで、圧縮した資料を付属する。
規模	約4万行の発話	約1万行の発話	約1万行の発話
コーパス化	整形化の必要あり	整形化の必要あり	整形化の必要あり

区分	日本語語彙大系	日本語動詞の結合価	分類語彙表
内容と規模	30万語の収録語は3000種の意味分類を用いて定義されている、大規模の日本語シソーラス。3000種の意味分類を用いて日本語の文型を定義している。	15万文規模の日本語文集合をもとに、「動詞がとりうる格と、その格を担う体言」について整備した、大規模な言語データ集。	延べ約９万６千語の日本語の単語を意味的に分類したもの
入手	有料（無料化予定あり）	有料	有料
検索ソフト	パッケージとして同梱	パッケージとして同梱	ない(PDFの索引のみ)
元の電子データを直接閲覧	基本的には不可能	基本的には不可能	可能(PDFファイル)
製作元	日本電信電話株式会社NTTコミュニケーション科学基礎研究所	日本システムアプリケーション荻野孝野・小林正博・井佐原均 編著	国立国語研究所

どうやって用例を集めるのか



文集合（生コーパス）
形態素解析
解析済みテキスト
＋
辞書
格納
検索
関係データベースシステム (e.g., MySQL)

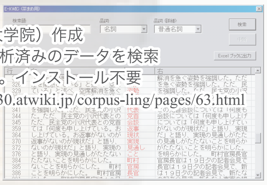
データ収集の流れ

● **形態素解析**
- 対象言語の文法の知識（文法のルールの集まり）や辞書（品詞等の情報付きの単語リスト）を情報源として用い、自然言語で書かれた文を形態素（Morpheme；言語で意味を持つ最小単位）の列に分割し、それぞれの品詞を判別する作業を指す。
- Chasen, Juman, Mecabなど
- 茶まめを使えば、Windowsでも簡単に利用可能。文脈付き用例抽出ツールを使ってより、高度な検索＆データ抽出ができる。

● **茶まめ**
- 国研日本語コーパス電子辞書班
- 形態素解析補助ツール
- 形態素解析は茶筌かMecab
- 入手方法：http://www.tokuteicorpus.jp/

● **KH Coder**
- 樋口耕一（立命館大学）作成
- 計量社会学ツール
- 分析機能やデータ抽出機能が充実
- 入手方法： http://khc.sourceforge.net/

● **E-KWIC**
- 淺尾 仁彦（京都大学大学院）作成
- 茶まめによる形態素解析済みのデータを検索
- エクセル上で動作する。インストール不要
- 入手方法： http://www30.atwiki.jp/corpus-ling/pages/63.html



どう分析するのか

Excelでできるもの

● **エクセルでデータ処理**
- テキストデータを処理する関数を覚える
  - ・例えば、left, right,mid,findなど
  - ・Excel(エクセル)基本講座：Excel(エクセル)関数一覧表 (http://www.eurus.dti.ne.jp/~yoneyama/Excel/kansu/itiran.html)

● **表層の頻度を利用した分析**
- 分布を調べる⇒ピボットテーブルを使う
- グループ間の類似度を調べる⇒相関係数を求める
- 語と語の共起強さを調べる⇒t-スコア, MIスコアを求める
- 分布の差を計算する⇒差異係数＆カイ二乗検定を行う

JMP, SPSSできるもの

● **エクセル+αでコーパス分析**
- エクセルでデータの下処理をしておく
- 主成分分析で情報を縮約する
- クラスタ分析でグループ化する
- 分散分析でグループを比較する

● **関連研究を見る**
- **コーパス日本語学**文献リスト：http://www30.atwiki.jp/corpus-ling

中本敬子・李在鎬・黒田航（編著）
『新しい認知言語学研究法入門』
心理実験、コーパス、作成による言語研究
ひつじ書房から2009.9に出版予定