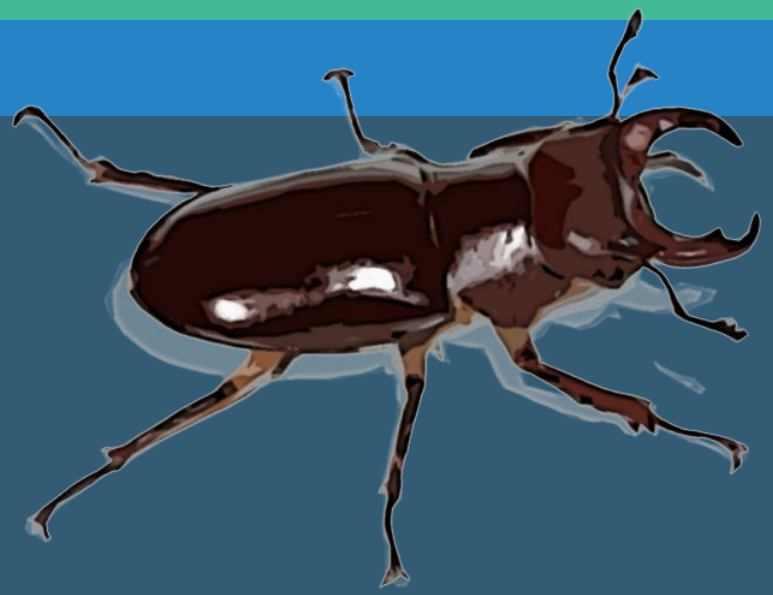
母語話者と学習者の話し言葉データを使ったコーパスシステム『Co-Chu』の研究事例

小森早江子　　ラニガン・マシュー

AATJ　Toronto 2017/3/16



CHUBU UNIVERSITY
AICHI JAPAN

本発表の構成

- コーパスシステム『Co-Chu』とは？
- MIスコアと記号について

中部大学コーパスシステム

Co-Chu【甲虫】

- コーパス構築
- コーパス分析

コーパス構築 ＋ コーパス分析

- なぜCo-Chuを開発したか
  - コーパス構築が難しい
  - 構築が終わっても分析も困難
  - ツールが多いが、強力なツールが少ない
  - 技術的な知識が多く必要とされる
- Co-Chuの特徴
  - 全機能のインタフェース統一
  - わかりやすい

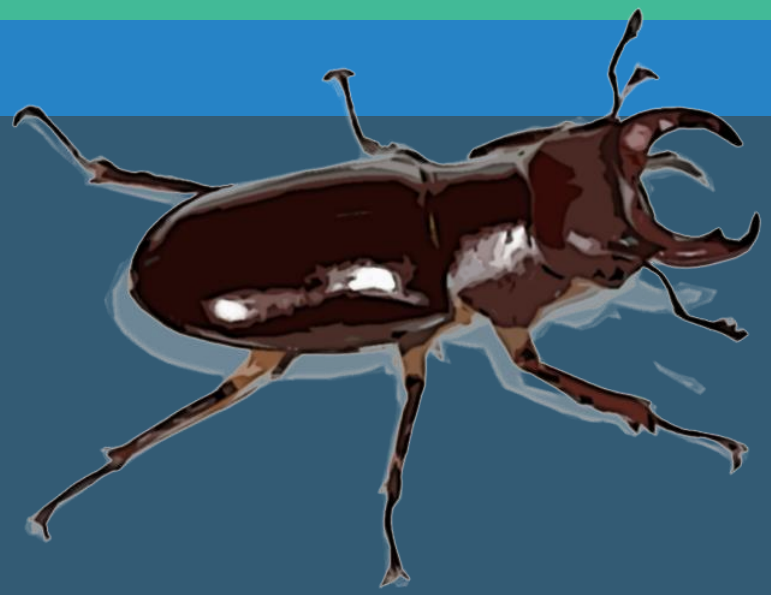
コーパス構築 ＋ コーパス分析

- コーパス構築の際に役立つ様々なツール
  - サブコーパス階層構造作成
  - データ取り込み
  - 自動的データ処理
    - 現在、形態素解析とコロケーション解析
    - 今後さらに追加したい（長単位形態素解析、係り受け解析など）
  - エラー処理

コーパス構築 ＋ コーパス分析

- 現在の分析機能の概要
  - サンプル検索（形態素連鎖の詳細検索）
  - サンプル頻度検索（結果の頻度表作成）
  - コロケーション検索

Co-Chuのインタフェース



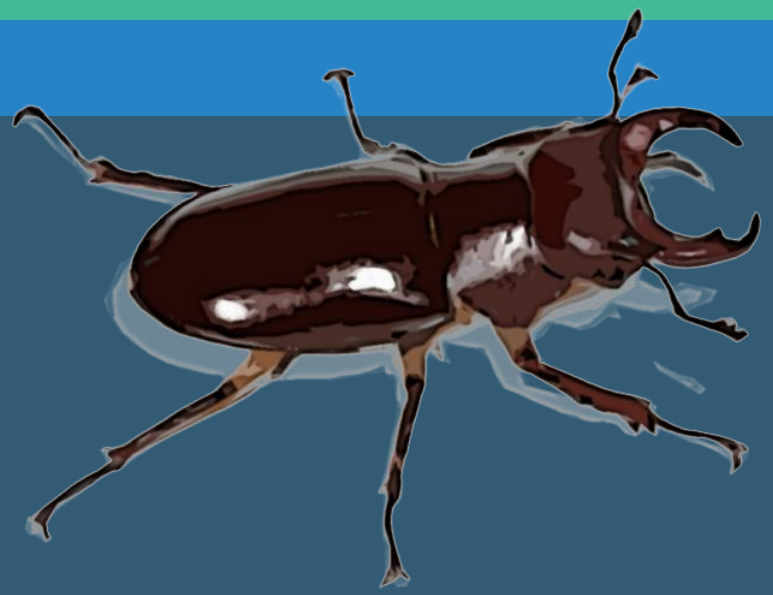
CHUBU UNIVERSITY
AICHI JAPAN

Co-Chuのインタフェース

- モードナビ
  - モード切り替え
- データナビ
  - 機能あるいはデータを選択する
- 機能エリア
  - コーパス情報、モードの主な機能など

Co-Chuの機能　①Build

コーパス構造作成機能

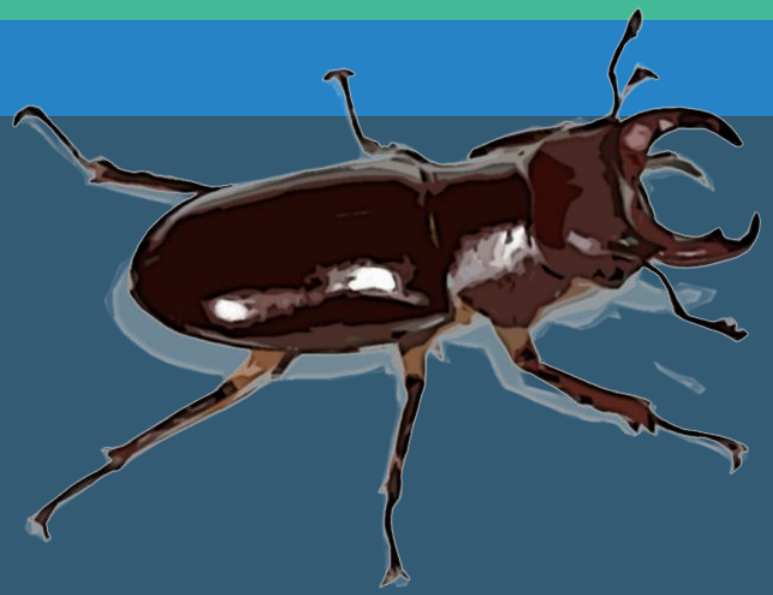


① Build

- コーパスツリー構造作成
- コーパスメタデータ・発話者データ編集

Co-Chuの機能　②Import

データ取り込み機能

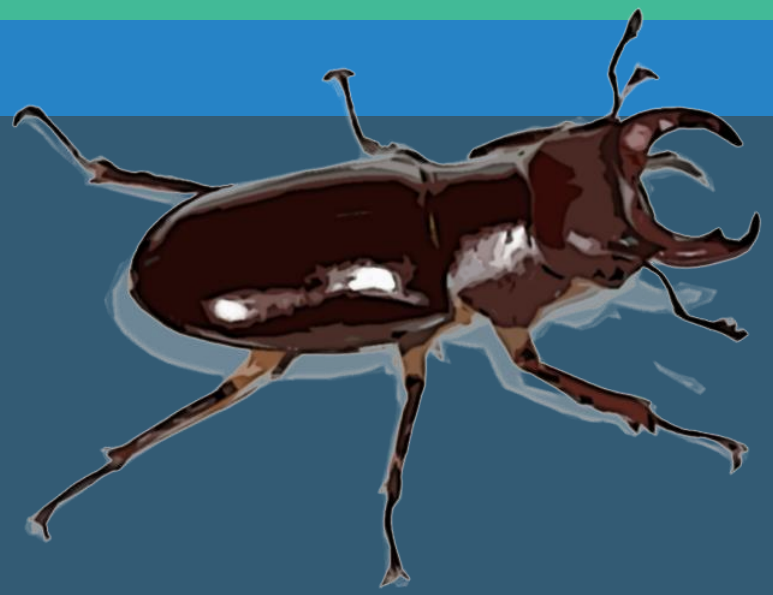


② Import

- データ取り込み
- 2種類
  1. 直接取り込み
     - 全ファイルを選択したコーパスに取り込む
  2. サブコーパス取り込み
     - 各ファイルをサブコーパスとして取り込む

Co-Chuの機能　③ Edit

データ編集・処理機能



③ Edit

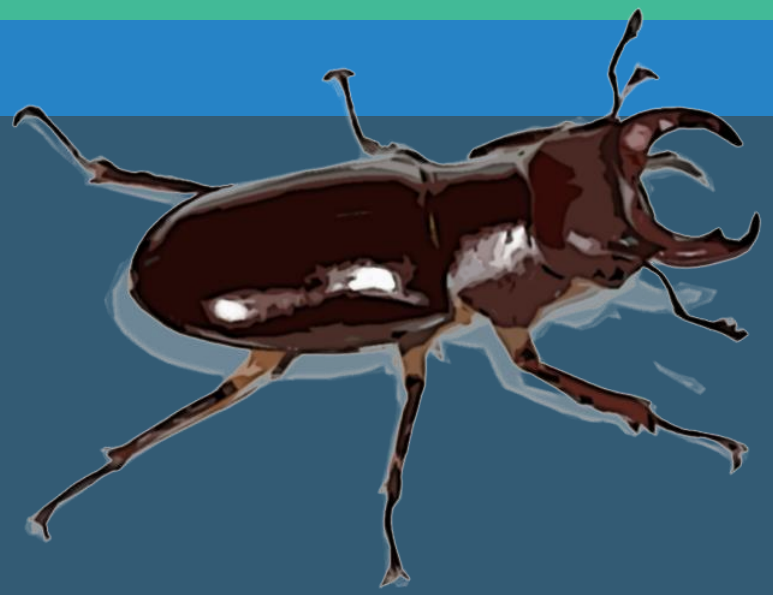
- データ編集
  - 例えば、不要のデータ
- 自動的データ処理
  - 形態素解析、コロケーション解析
- エラー処理
  - Nbest候補選択

③ Edit

- 基本的に
  1. データ編集
  2. 形態素解析
  3. エラー処理
  4. コロケーション解析など

Co-Chuの機能　④ Analyze

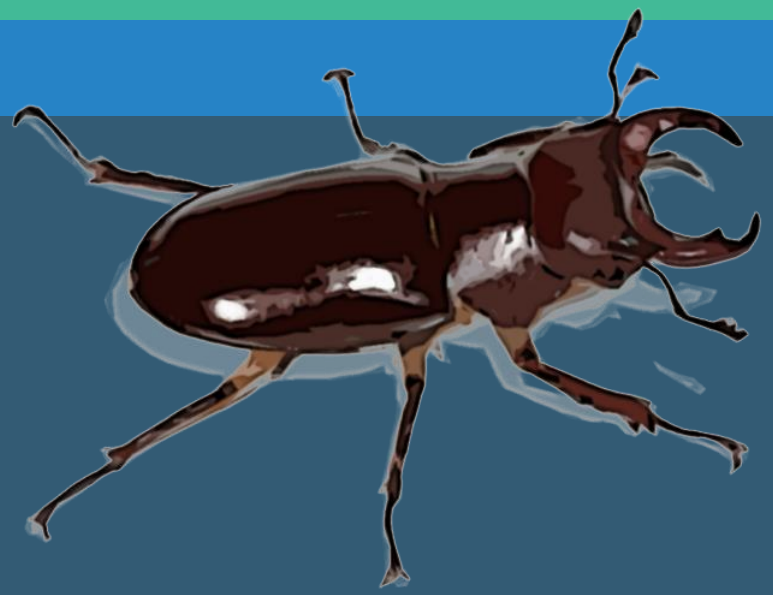
コーパス分析機能



④ Analyze

- 対象データセット選定
  - 何を対象にするか
  - メタデータ・話者フィルタ
- 検索項目定義
  - 形態素の連鎖
- アウトプット設定
  - 検索モードによって、様々な設定ができる

Co-Chuのまとめ



コーパス開発が困難

- 技術的な知識が必要
- ツールが多い
- 強力ツールが少ない

特に話し言葉が困難

- データが処理しにくい

コーパスシステム『Co-Chu』

コーパスシステム『Co-Chu』

同じインタフェース

コーパス構築

- コーパスツリー構造作成
- データ取り込み
- データ編集・処理

＋

コーパス分析

- 語を探す
- 語を数える
- Nグラムを見る

2　MIスコアと記号

I. MIスコア
II. Co-Chu雑談コーパスデータの例
III. 記号の処理について

I. MIスコアとは

MIスコアとは、コロケーションの強度を表す指標

- 数値が大きければ大きいほどコロケーションが強い

（通常、MIは3.0以上が興味深い, Church and　Hanks,1990）

- コロケーション分析の手法を見出し、話し言葉や学習者のデータの分析に用いる

MIスコアの計算式

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Church and Hanks (1990)

ある語xが特定の語yと共起する確立を試算する

Wei and Li (2013)の提案する新しい方法

- 前後する2語の関係だけでなく、３語以上の共起関係も調べる
- 同じような表現を絞り込む方法

2-gram (w_1, w_2)	$w_1 * w_2$
3-gram (w_1, w_2, w_3)	$w_1 * w_2, w_3$
	$w_1, w_2 * w_3$
4-gram (w_1, w_2, w_3, w_4)	$w_1 * w_2, w_3, w_4$
	$w_1, w_2 * w_3, w_4$
	$w_1, w_2, w_3 * w_4$
5-gram $(w_1, w_2, w_3, w_4, w_5)$	$w_1 * w_2, w_3, w_4, w_5$
	$w_1, w_2 * w_3, w_4, w_5$
	$w_1, w_2, w_3 * w_4, w_5$
	$w_1, w_2, w_3, w_4 * w_5$

MIスコアのピークで絞り込む方法 (Wei and Li 2013より）

Table 2. Treatment of repeated *n-grams* of different sizes

n-grams	Frequency	New-MI	
a metropolitan	29	3.3032	☒
a metropolitan area	27	4.9413	☑
a metropolitan area on	3	4.4908	☒
a metropolitan area on population	3	5.6129	☒
a metropolitan area on population change	3	5.6327	☑

上から順にNew-Miのスコアを見ると、3.30 ＜ 4.94 ＞ 4.49 ＜ 5.61 ＜ 5 .63 となる

II　Co-Chu雑談コーパス

- オリジナルデータ　少しずつ収集し、データ化進行中

- NSとNNSとの１対１の発話
- テーマ「最近あったことについて」
- 中国人上級日本語学習者CS　4人と日本人
  30分 x 12発話
- 韓国人上級日本語学習者KS　2人と日本人
  30分 x ６発話

「思う」の後にくる表現の例

Nグラム表現	N	頻度	MIスコア
思っ　た　の　だ	4	3	3.158922949
思っ　て　た	3	15	3.861233376
思っ　て　た　ん　だ	5	4	4.834194068
思っ　て　た　ん　だ　けど	6	4	7.143562697
思っ　てる　けれど	3	3	3.724219712
思わ　れ　てる	3	3	4.572438284
思う　ん　だ　けれど	4	8	3.886156173
思う　ん　だ　わ	4	3	6.071548393

発話例：

日本語学習者(CS-O）と母語話者（JS-H）の会話データ

話者	発話ターン
【CS-O】	あのー、なんか、んー、あのー、なんか、|何|(なん)だろうね、
【CS-O】	病院↑、病院の中の
【JS-H】	うん、そうそう。／／
【CS-O】	／／|何|(なに)これ
【JS-H】	心電図だ、心電図。
【CS-O】	ああ、心電図。／／そうだよね｛笑｝。

III　記号の処理

- 書き起しデータには次のようなものが含まれている
  - ↑　：　上昇イントネーション
  - 、。：　日本語として読みやすくする句読点
  - （　）：　漢字語彙の読みを表す
  - ／／：　オーバーラップを表す
  - {笑}：　非言語情報を示す

など

- これらを取り除いたもの→「記号なし」

Nグラム	N	あり頻度	ありMIスコア	なし頻度	なしMIスコア
そう　だ　ね	3	214	3.64214852	219	3.704140817
と　思う	2	182	4.09854427	197	3.700215816
日本　語	2	164	5.60122745	164	4.915798216
そう　そう　そう	3	94	3.7408931	104	4.859759581
日本　人	2	96	6.14553923	96	4.47694611
どう　する	2	54	3.90372685	54	3.253894957
たり　する	2	49	4.8434025	53	3.405570128
なっ　ちゃう	2	52	5.47285353	52	3.935307656
て　くれる	2	47	6.41763791	50	4.090774605
勉強　する	2	50	4.675512	50	3.97358256
か　も　知れ　ない	4	45	5.61040701	47	4.619948204
ああ　そう　か	3			45	3.123976626
一　月	2			44	4.97932846
そうそう　そうそう	2	28	4.0964367	44	5.249054069
もん　ね	2	41	4.63996878	41	3.666705265
確か　だ	2	41	3.74553863	41	3.123523187
御　金	2	41	8.39037568	41	6.70884479

- 同じ文字列でも「記号あり」と「記号なし」ではMIスコアが異なる。

↓

MIスコアを算出する際には、

- 記号の取り扱いに注意する。
- 調べたい語や状況などに合わせて、記号を処理する必要がある。

まとめ

- コーパスシステム「Co-Chu」について説明
  Co-Chuには、「コーパス構築」と「コーパス分析」の部分がある

- オリジナルデータCo-Chu雑談コーパスを使ってMIスコアを算出

- MIスコアの算出には記号の取り扱いに注意する必要がある

コーパスシステム「Co-Chu」は開発途中
今後も改良・分析を進めていきたい

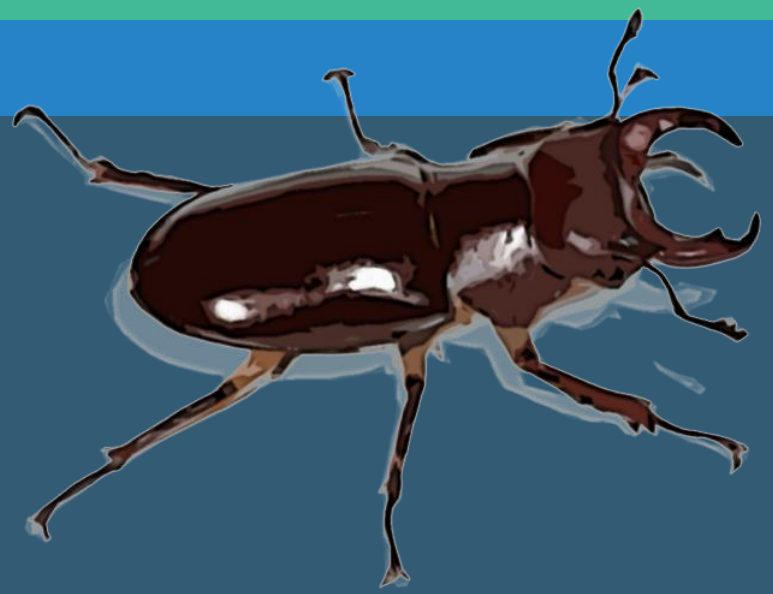
参考文献

- Church, K.W. and Hanks P.(1990) Word association norms, mutual information, and lexicography, *Computational linguistics* 16:1, 22-29.
- Wei, N. and Li, J. (2013) A new computing method for extracting contiguous phraseological sequences from academic text corpora, *International Journal of Corpus Linguistics* 18:4, 506–535.
- 小森早江子、山本裕子、本間妙、ラニガン　マシュー　(2015）「日本語教育・研究のためのテキスト分析システム” Co-Chu”」The Proceedings of CASTEL/J in Hawaii 2015.

ご清聴ありがとうございました。
Co-Chuプロジェクトメンバー

小森早江子
山本裕子
本間妙
ラニガン・マシュー

AATJ Toronto 2017/3/16



co-chu.org

CHUBU UNIVERSITY
AICHI JAPAN