日本語意味フレーム分析における対訳コーパスの利用

ワークショップ「コーパス利用とこれからの認知言語学：
用法基盤主義をカケ声でおわらせないためには, 何を, どうするべきか」

金丸 敏幸

<kanamaru@hi.h.kyoto-u.ac.jp>

京都大学大学院人間・環境学研究科

2005年09月18日 日本認知言語学会 第6回 全国大会

はじめに

- 何が目的？
  - コーパスを使った動詞「襲う」の意味フレーム分析
- 何をした？
  - 共起語の収集とその分類, 分析
- 何が分かった？
  - 頻度順リストだけではダメ
  - 意味順に並び替えると分析が楽

日本語意味フレーム構築の流れ

1. 日本語動詞の英語意味フレームを特定
2. 英語意味フレームの Lexical Units (LUs) を含む対訳文を収集
3. 得られた対訳文の中に高頻度で出現する語から日本語の LUs を決定
4. 英語意味フレームを元に, 日本語意味フレームを構築

日本語意味フレーム構築の流れ

1. 日本語動詞の英語意味フレームを特定
2. 英語意味フレームの Lexical Units (LUs) を含む対訳文を収集
3. **得られた対訳文の中に高頻度で出現する語から日本語の LUs を決定**
   - **→今回の発表では，この過程を説明.**

対訳文の収集

- 英語意味フレームの LUs を収集
  - フレーム記述にある, LUs (今回は v のみ)を各フレーム毎に収集した.
    - <Attack> フレームの例
      - ambush, assault, attack, charge, fall, invade, jump, lay, set, storm, strike
  - → 3フレーム合計で, 全93動詞を収集

- これらの単語を含む対訳文をフレーム毎に抽出. → 全982対を抽出(人手で修正).

LUs の決定

- 対訳文の日本語文を KH Coder で解析
  - 「品詞別 出現回数順リスト」で名詞の頻度順リストを得る.
- 名詞の頻度順リストの上位100 (かつ, 頻度3以上) を msort (村田他 2000) で, 意味別にソートする.

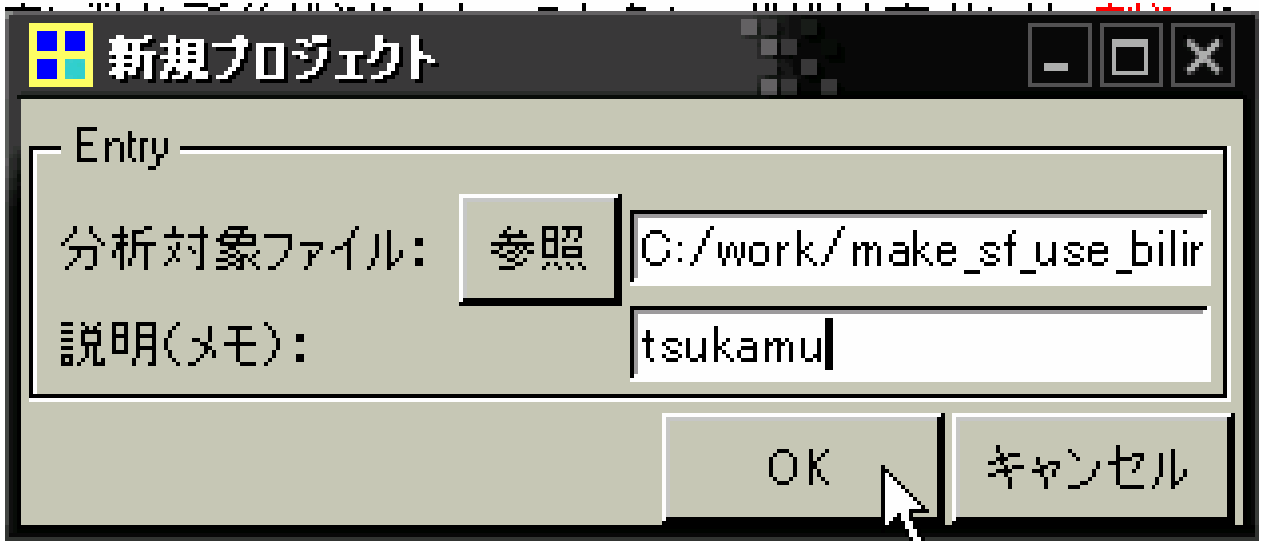
KH Coderの概略

- 樋口耕一氏 （日本学術振興会 PD, 関西学院大学 社会学部 非常勤講師, 京都学園大学 人間文化学部 非常勤講師）が作成.
- 入手先
  - http://khc.sourceforge.net/
- 特徴
  - テキストを定量分析するためのツール.
    →コーパス分析にも応用可能.

KH Coderの使い方

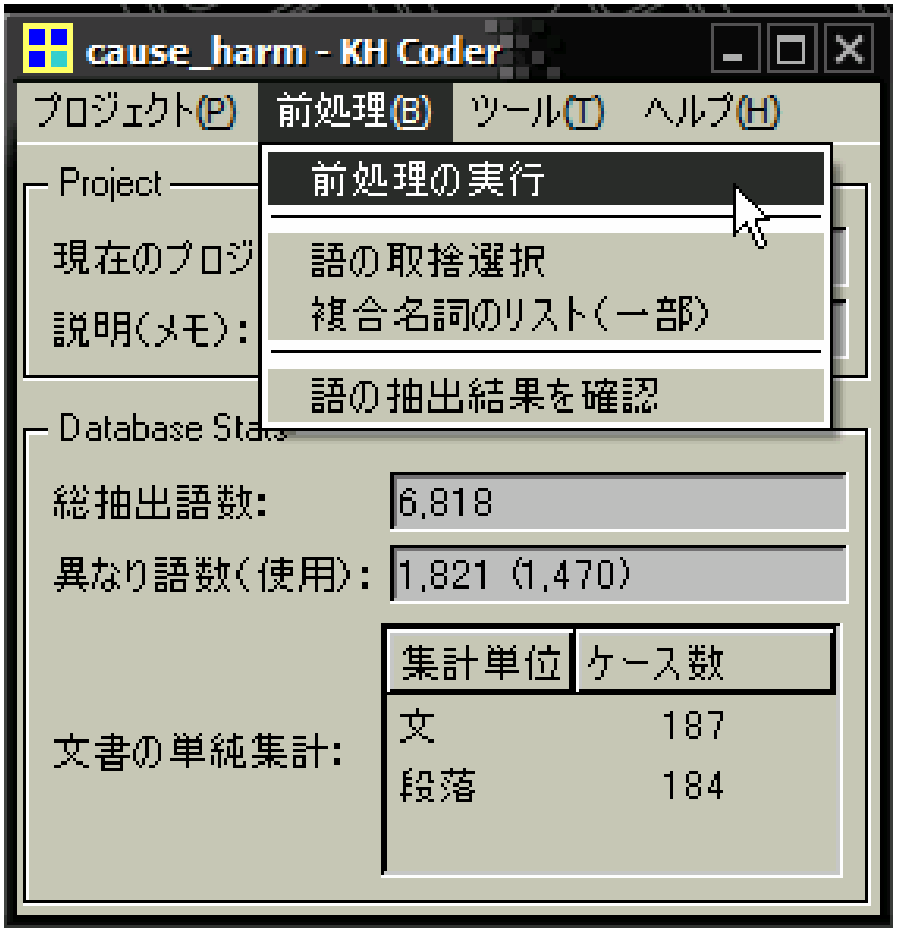
- 新規プロジェクトを作成
  - 今回の分析では，一つのフレーム毎に，対訳文が得られているので，それらをまとめて入力する．

KH Coder



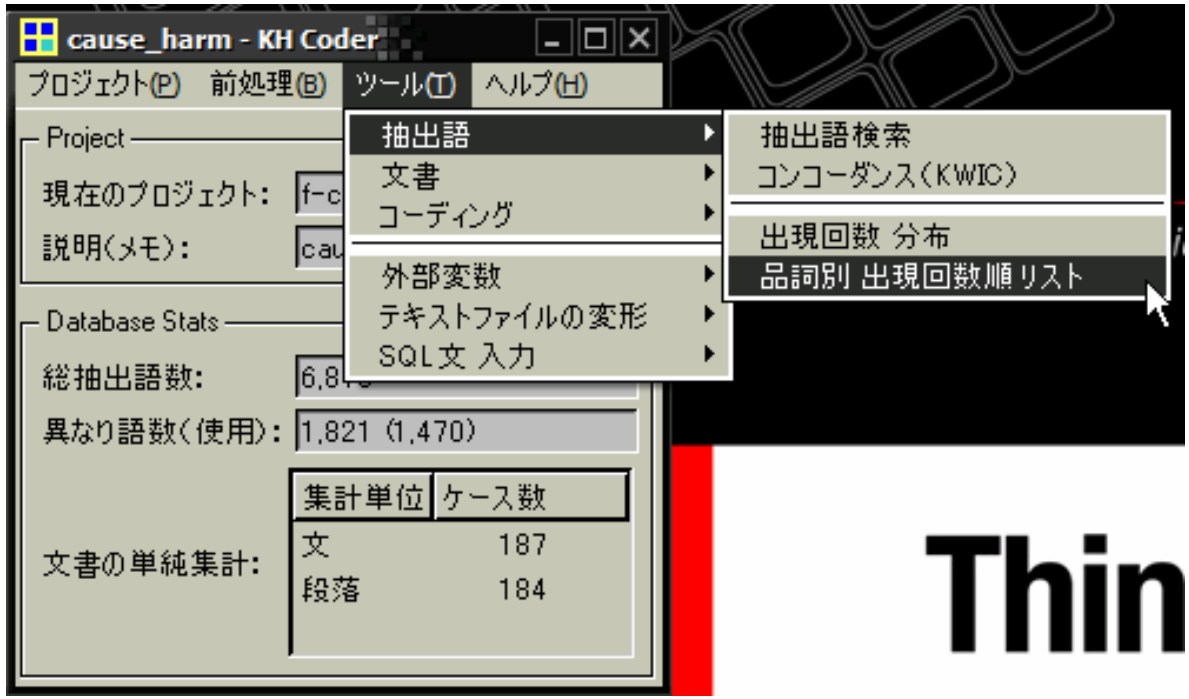
- 分析対象ファイルと説明を入力.

KH Coder



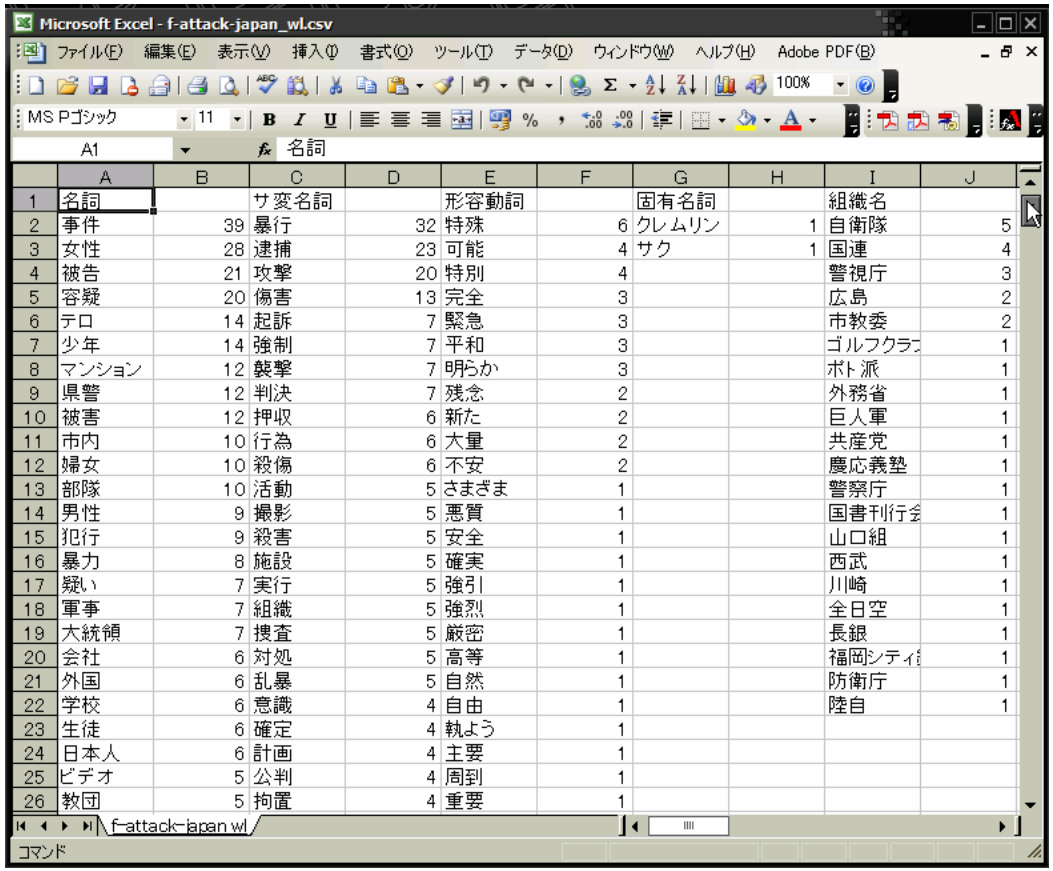
- 入力テキストを選択したら，前処理を実行．
- 前処理とは，chasen（松本他 1999）で，入力文に対し，形態素解析を行い，結果をデータベースに格納することを指す．

KH Coder



- 前処理が終了したら，メニューから「品詞別 出現回数順リスト」を選択.

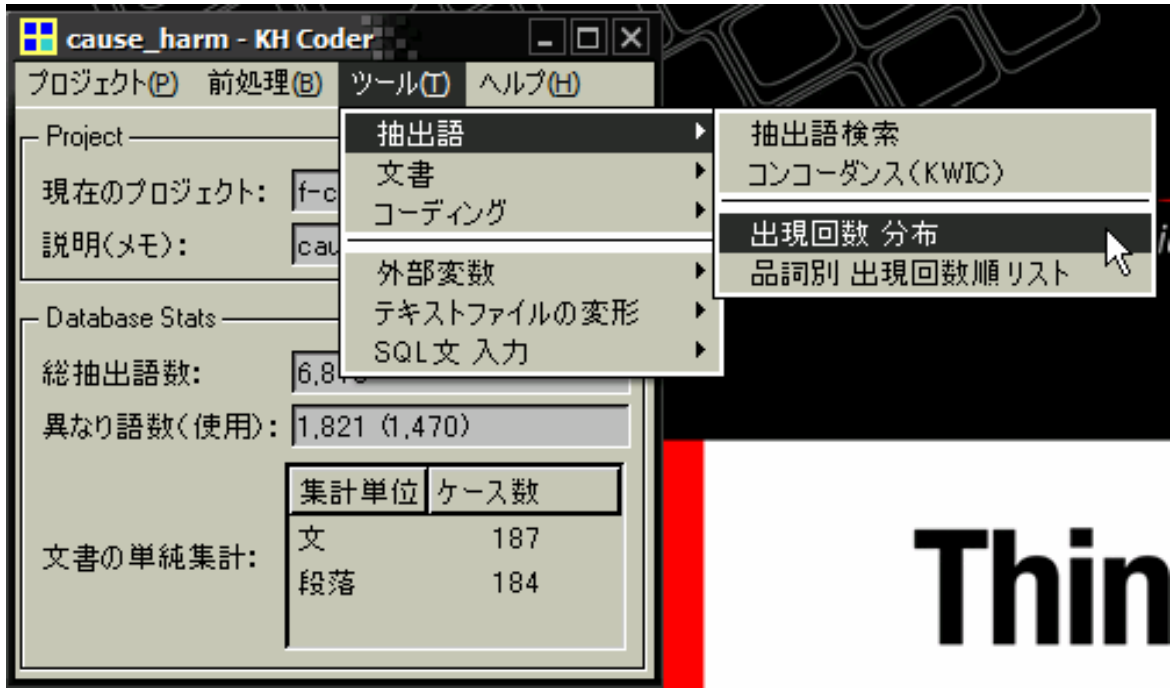
KH Coder



	A	B	C	D	E	F	G	H	I	J
1	名詞		サ変名詞		形容動詞		固有名詞		組織名	
2	事件	39	暴行	32	特殊	6	クレムリン	1	自衛隊	5
3	女性	28	逮捕	23	可能	4	サク	1	国連	4
4	被告	21	攻撃	20	特別	4			警視庁	3
5	容疑	20	傷害	13	完全	3			広島	2
6	テロ	14	起訴	7	緊急	3			市教委	2
7	少年	14	強制	7	平和	3			ゴルフクラ	1
8	マンション	12	襲撃	7	明らか	3			ボト派	1
9	県警	12	判決	7	残念	2			外務省	1
10	被害	12	押収	6	新た	2			巨人軍	1
11	市内	10	行為	6	大量	2			共産党	1
12	婦女	10	殺傷	6	不安	2			慶応義塾	1
13	部隊	10	活動	5	さまざま	1			警察庁	1
14	男性	9	撮影	5	悪質	1			国書刊行会	1
15	犯行	9	殺害	5	安全	1			山口組	1
16	暴力	8	施設	5	確実	1			西武	1
17	疑い	7	実行	5	強引	1			川崎	1
18	軍事	7	組織	5	強烈	1			全日空	1
19	大統領	7	捜査	5	厳密	1			長銀	1
20	会社	6	対処	5	高等	1			福岡シティ	1
21	外国	6	乱暴	5	自然	1			防衛庁	1
22	学校	6	意識	4	自由	1			陸自	1
23	生徒	6	確定	4	執よう	1				
24	日本人	6	計画	4	主要	1				
25	ビデオ	5	公判	4	周到	1				
26	教団	5	拘置	4	重要	1				

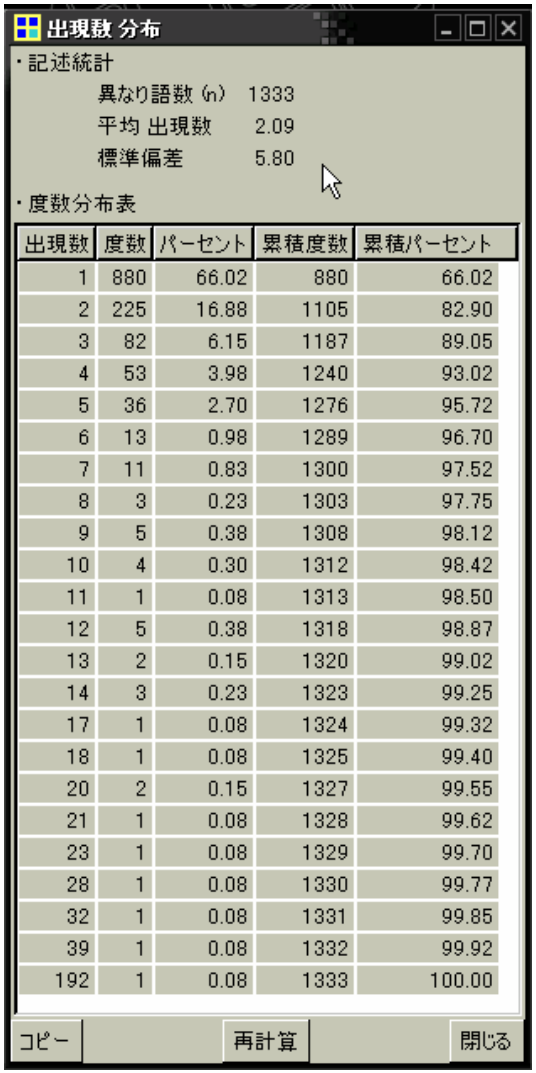
- （あれば）Excel が起動し，品詞別頻度順リストを表示してくれる．
- 品詞は，前処理で使用した chasen が判断したもの．

KH Coder



- 出現回数は，読み込んだ文の数などによって，左右されるので，全体に対する出現割合を調べる必要がある．

KH Coder


出現数 分布

・記述統計

異なり語数 (n)　1333
平均 出現数　2.09
標準偏差　5.80

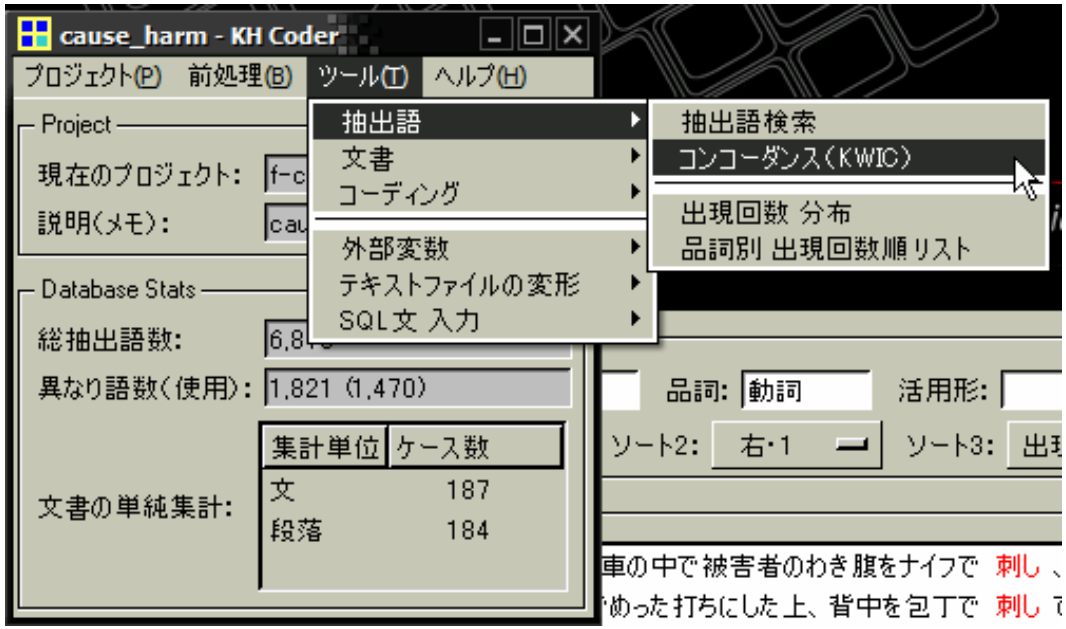
・度数分布表

出現数	度数	パーセント	累積度数	累積パーセント
1	880	66.02	880	66.02
2	225	16.88	1105	82.90
3	82	6.15	1187	89.05
4	53	3.98	1240	93.02
5	36	2.70	1276	95.72
6	13	0.98	1289	96.70
7	11	0.83	1300	97.52
8	3	0.23	1303	97.75
9	5	0.38	1308	98.12
10	4	0.30	1312	98.42
11	1	0.08	1313	98.50
12	5	0.38	1318	98.87
13	2	0.15	1320	99.02
14	3	0.23	1323	99.25
17	1	0.08	1324	99.32
18	1	0.08	1325	99.40
20	2	0.15	1327	99.55
21	1	0.08	1328	99.62
23	1	0.08	1329	99.70
28	1	0.08	1330	99.77
32	1	0.08	1331	99.85
39	1	0.08	1332	99.92
192	1	0.08	1333	100.00

コピー　再計算　閉じる

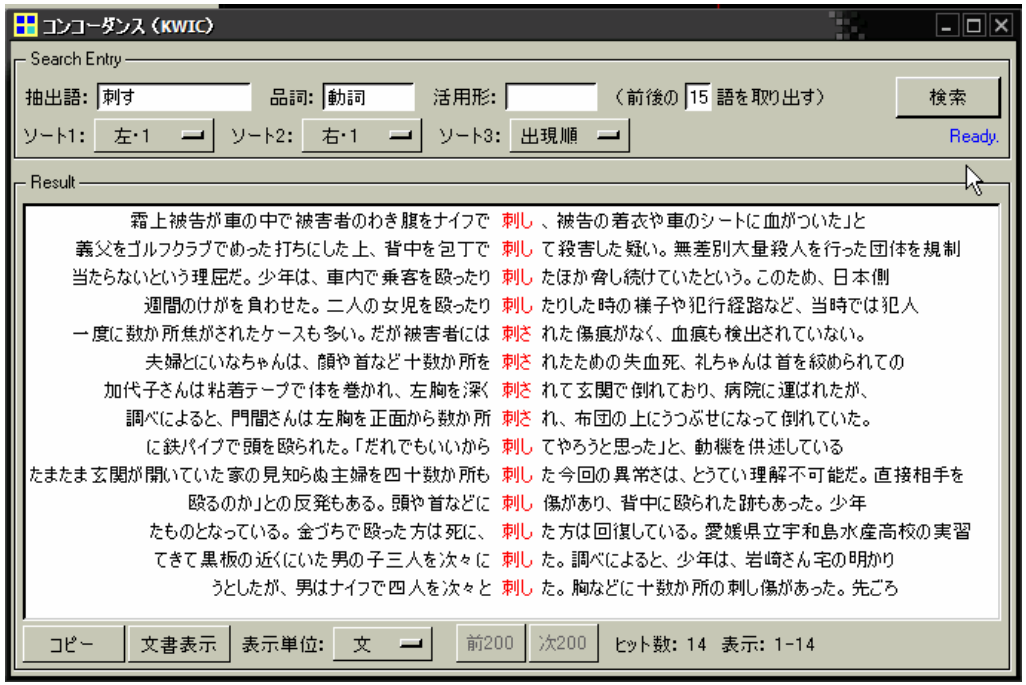
- フレームの特徴を抽出するには，頻度は比較的高い方がよい．
- 文章中から，有意に特徴的な単語を抽出する方法には，いくつかの手法がある（今回は割愛）．

KH Coder



- 単語の文脈中での用法も調べることができる.
  - KWIC (KeyWord In Context) 表示も可能.

KH Coder



- 形態素解析済みであれば，表層形に関係なく，検索，表示が可能.

得られた名詞

<Attack>フレームの例 (上位30例)

名詞	頻度	名詞	頻度	名詞	頻度
事件	39	被害	12	大統領	7
暴行	32	県警	12	襲撃	7
女性	28	マンション	12	軍事	7
逮捕	23	部隊	10	強制	7
被告	21	婦女	10	疑い	7
容疑	20	市内	10	起訴	7
攻撃	20	犯行	9	日本人	6
少年	14	男性	9	生徒	6
テロ	14	暴力	8	殺傷	6
傷害	13	判決	7	行為	6

頻度順表示での限界

- あるフレームの下に集められたデータ内における高頻度語は，確かにフレームに関係していそうな感じがする．
  1. 事件 39 … 「襲う」と事件になる？
  2. 暴行 32 … 暴行する？ 暴行事件？
- 頻度だけでは，フレームと語の関係が分かりにくい．
- →頻度だけを調べるだけでは，不十分．

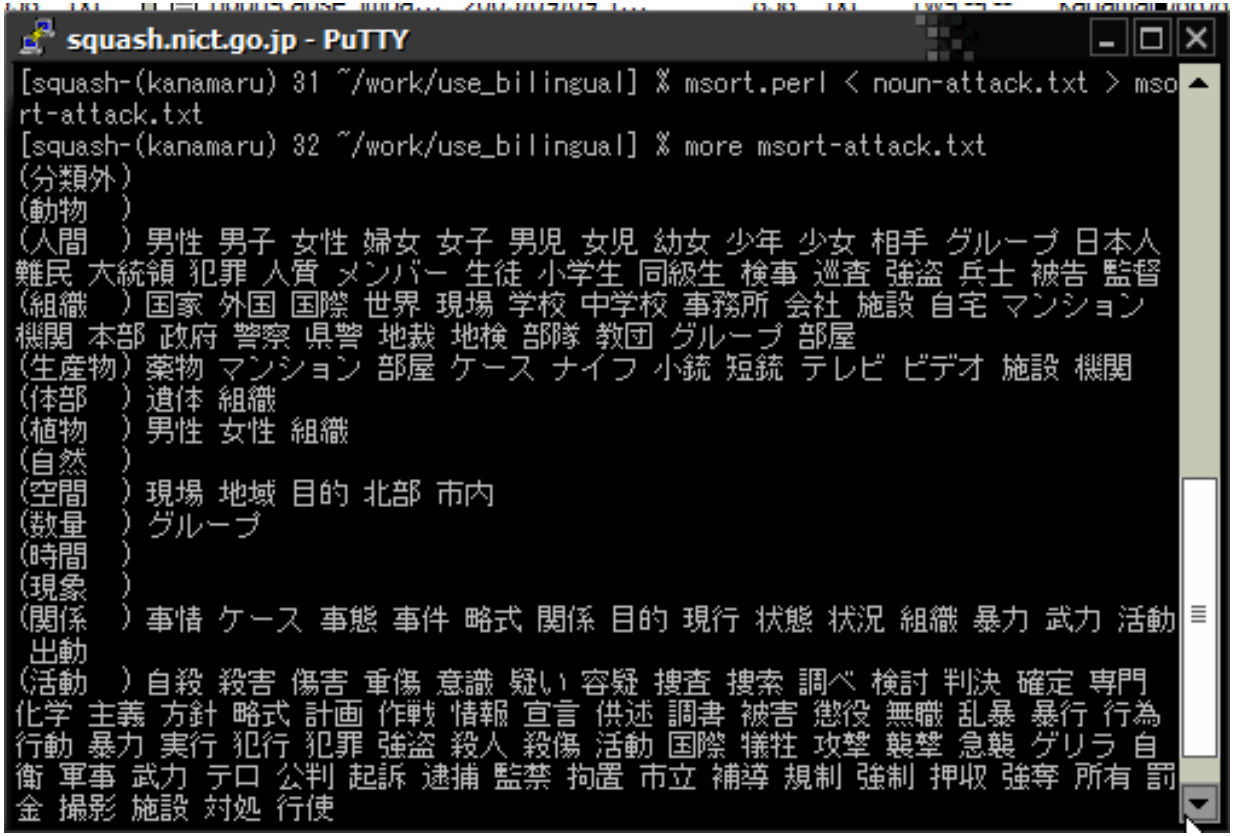
頻度順表示での限界

- 得られた語にどのような性質があるのか，どのような文脈で生起しているかを分析する必要性がある．
  - 語の「分類」を行う必要がある．
    - → 意味ソート(msort) を使用した．
  - 語が生起している文脈を調査する必要がある．
    - → 実際のコーパスを分析した．

意味ソート

- 村田真樹氏（情報通信研究機構）が作成.
  - 現在は, 一般公開されていない.
  - 使用してみたい方は, 直接ご連絡を.
    - E-mail : murata@nict.go.jp
- 特徴
  - 単語を分類語彙表（国立国語研究所）の分類IDに従って, 並べ替える.
  - 分類語彙表にない単語でも, ある程度類推して, 並び替えてくれる.

意味ソート



- コマンド入力で使用する.

意味ソートの結果

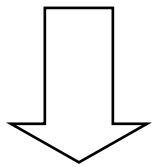
- (動物)
- (人間) 男性 男子 女性 婦女 女子 男児 女児 幼女 少年 少女 相手 グループ 日本人 難民 大統領 犯罪 人質 メンバー 生徒 小学生 同級生 検事 巡査 強盗 兵士 被告 監督
- (組織) 国家 外国 国際 世界 現場 学校 中学校 事務所 会社 施設 自宅 マンション 機関 本部 政府 警察 県警 地裁 地検 部隊 教団 グループ 部屋
- (生産物) 薬物 マンション 部屋 ケース ナイフ 小銃 短銃 テレビ ビデオ 施設 機関
- (体部) 遺体 組織
- (植物) 男性 女性 組織
- (自然)
- (空間) 現場 地域 目的 北部 市内
- (数量) グループ
- (時間)
- (現象)
- (関係) 事情 ケース 事態 事件 略式 関係 目的 現行 状態 状況 組織 暴力 武力 活動 出動
- (活動) 自殺 殺害 傷害 重傷 意識 疑い 容疑 捜査 捜索 調べ 検討 判決 確定 専門 化学 主義 方針 略式 計画 作戦 情報 宣言 供述 調書 被害 懲役 無職 乱暴 暴行 行為 行動 暴力 実行 犯行 犯罪 強盗 殺人 殺傷 活動 国際 犠牲 攻撃 襲撃 急襲 ゲリラ 自衛 軍事 武力 テロ 公判 起訴 逮捕 監禁 拘置 市立 補導 規制 強制 押収 強奪 所有 罰金 撮影 施設 対処 行使
- (その他)
- (未定義) シー 睡眠薬 組員 同署 特捜

出現語彙の分析

- 意味フレームの要素になるもの.
  - (人間) 男性 男子 女性 婦女 … 巡査 強盗 兵士
  - ただし, 分類項目内の全てが同じ要素になるとは限らない.
- 意味フレームの要素にならないもの.
  - (活動) … 殺害 傷害 重傷 容疑 捜査 捜索 供述調書 被害 懲役 乱暴 暴行 暴力 犯行 犯罪 強盗 殺人 殺傷 …
  - →「襲う」フレームに間接的に関わる語.

コーパス内での実例

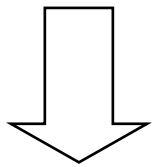
- これまで拷問などに加わったハイチ　兵士 らが大統領支持派や米軍兵士らへのテロ活動を行う恐れもある。



- 兵士 が 兵士 へ テロ活動を行う.
- → 兵士<襲い手>, 兵士<襲われ手>
- → テロ活動を行う → <襲う>

コーパス内での実例

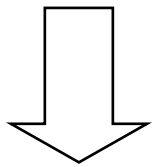
- 確定判決によると、佐藤死刑囚は五九年に山口市内で七歳の幼女を　　殺害 して無期懲役が確定し、…



- 死刑囚 が 幼女 を殺害した.
- → 死刑囚<襲い手>, 幼女<襲われ手>
- → 殺害する<殺害> ← <襲う>の事態の結果.

コーパス内での実例

- 三月にはバングラデシュ部隊駐屯地が襲撃 されて死者を出し、先週もポト派の襲撃でブルガリア兵六人が死傷し…



- (ゲリラ兵が) 駐屯地 を襲撃した.
- → (ゲリラ兵<襲い手>), 駐屯地<襲われ手>
- → 襲撃する<襲撃> ＝ <襲う>.

構築できた日本語意味フレーム

- <Attack> FEs:

  Core:

  - Assailant: 男性 強盗 兵士 被告 部隊 教団 グループ メンバー
  - Victim: 男子 女性 婦女 女子 男児 女児 幼女 少年 少女 相手 日本人 難民 大統領 人質 生徒 小学生 同級生 監督 世界 学校 中学校 事務所 会社 施設 自宅 マンション …

  Non Core:

  - Place: 現場 地域 目的 北部 市内
  - Weapon: ナイフ 小銃 短銃

<襲う>と関わる名詞群

- 「襲う」行為の名称
  - 乱暴 暴行 暴力 実行 襲撃 攻撃
- 「襲う」事態の結果に対する名称
  - 殺害 傷害 強盗 殺人 殺傷
- 「襲う」行為の結果
  - 重傷
- 「襲う」行為の結果生じる「犯罪」フレームや「裁判」フレームと関わる語彙
  - 疑い 容疑 捜査 捜索 調べ 検討 判決 供述 調書 懲役 公判 起訴 逮捕 監禁 拘置 罰金 犯行 犯罪 …

用法基盤モデルとの関わり

- 実際の用例との対応を見る.
  - 理論先行形の議論はしない.
  - 理論的整合性よりも, 実際の使用を優先.
- 高頻度語は意味フレームと強い関わりを持つ.
  - (Type)頻度効果が現れている.
  - 語とフレームだけではなく, フレーム同士でも頻度効果は確認できる.
  - ただし, 関わり方はそれぞれの語で異なる.

まとめ

- 何が目的？
  - コーパスを使った意味フレーム分析
- 何をした？
  - 共起語の収集とその分類, 分析
- 何が分かった？
  - 頻度順リストだけではダメ
  - 意味順に並び替えると分析が楽
  - 自動化できるところは, 自動化した方がよい.

参考文献

- C. J. Fillmore, C. R. Johnson, and M. R. L. Pentruck, “Background to FrameNet,” International Journal of Lexicography, Vol.16, No.3, pp.235-250, 2003.
- 黒田航, 中本敬子, 金丸敏幸, 龍岡昌弘, 野澤元.「意味フレーム」に基づく概念分析の射程: Berkeley FrameNet and Beyond. 日本認知言語学会第5回大会Conference Handbook, pp. 133-153. 日本認知言語学会(JCLA), 2004.
- 村田真樹, 神崎享子, 内元清貴, 馬青, 井佐原均. 「意味ソートmsort－－意味的並べかえ手法による辞書の構築例とタグづきコーパスの作成例と情報提示システム例－－」, 自然言語処理, Vol.7, No.1, pp.51-66, 2000.
- 中本敬子, 黒田航, 野澤元. 「素性を利用した文の意味の心内表現の探索法」, 認知心理学研究. 印刷中.
- 佐藤弘明. 「英語語彙データベース FrameNet 検索用ソフトウェア FrameSQL」, 情報学研究, Vol.25. pp.1-14, 専修大学, 2005.
- 内山将夫, 井佐原均. 「日英新聞記事および文を対応付けるための高信頼性尺度」, 自然言語処理, No.10, Vol.4, pp.201-220, 2003.
- 松本裕治, 北内啓, 山下達雄, 平野喜隆, 松田寛, 浅原正幸. 「日本語形態素解析システム 『茶筌』 Version 2.0 使用説明書 第二版」NAIST Technical Report NAIST-IS-TR, 1999.

謝辞

- 独立行政法人 情報通信研究機構(NICT)
  - 村田真樹 氏
- 各種ツール・データを作成されている皆様
  - 茶筅(chasen) … 奈良先端大 松本研究室
  - KH Coder … 樋口耕一 氏
  - 対訳コーパス … 内山将夫 氏(NICT)
  - BFNプロジェクトの皆様
  - FOCALプロジェクトの皆様

謝辞

ご静聴ありがとうございました.