Corrected Version

# コーパス事例の観察に基づく日本語清掃表現の記述的一般化, 並びに自作例による妥当性の検討

黒田 航

NICT けいはんな研究所

第26英語学会ワークショップ

2008/11/15, 筑波大学

# 本発表の狙い

- 日本語コーパス調査から得られた記述的一般化の一部を作例を通じて検証する
  - Word Sketch Engine (Kilgarriff & Tugwell 01; Srdanovic, *et al.* 08) の使い方の簡単な解説
- 外観ベースの調査法と内観ベースの調査法の使い分けのし方の提示

# 実例と作例のうまい使い分け

# 不毛なイガミ合い

- 内観至上主義者 (生成言語学者に多い) の主張
  - I-Language の研究にコーパスなんて要らない
  - Grammar is Grammar and Usage is Usage (Newmeyer 03)

# 不毛なイガミ合い

- 外観至上主義者(is-a 客観性至上主義者) (コーパス言語学者に多い) の主張
  - 直観なんて (客観性がないから) アテにならン
  - 作例は信用できン．実例しか信用したらアカン

# どっちが正しいの??

# “中道” を行くために

- 内観法で得られるデータ/証拠と外観法で得られるデータ/証拠には，いずれにも長所と短所ある
  - “内観法が常に外観法に優先する” (内観至上主義者の主張) と
  - “外観法が常に内観法に優先する” (外観至上主義者の主張) とはどっちも誤り
- とすれば，時と場合に応じて二つを使い分けるのがスジ

# 答えの自明ではない問い

- 内観法と外観法の有効な使い分けとは??
- 作例と実例の有効な使い分けとは??

# 実例と作例の関係

定量的分析と定性的分析の統合

# “実例” とは何か?

- 科学的な文脈で問題になる “実例” 性とは，検討される例が事実に対して代表性をもっているかどうか
  - 実証的な意味での “実例” とは (説明とは独立に) “実際に使われた事例” ということ
- 重要な点
  - 代表性をもたない例は本当の意味での “実例” ではない
  - 作例の多くは，実証的な意味での “実例” ではない

# “実例” はどう集めるべきか?

- この意味での “実例” の収集に作例は不向き
  - 人の想像力には限界があ (り，想起力には抑制がかかってい) るので，作例ベースの “実例” の収集は常に被覆率が不足する
  - その上，研究者の想像力と想起力には確証バイアスがかかっている
  - 結果として，作例で得られるデータには常に偏りがあり，代表性を欠いたものになる

# “実例” はどこにあるか?

- 一面では，実例はありとあらゆるところにあるが，それが資料として整備されているかとなると話は別
- 十分な代表性をもつ実例の集合を資料として利用するとなると，現状では規模の大きな電子化均衡コーパスか Web データ以上のものはない
  - 後者に関しては微妙な著作権問題が存在するので注意が必要

# “実例” だけで十分か?

- 実は，実証的研究には実例 (examples) だけでなく反実例 (*anti*examples) も必要
  - 記述的一般化や理論的予測の評価には “正例” だけでなく “負例” も必要
    - 汚れを洗い落とす (実例の例) vs. *服を洗い落とす (反実例の例)
  - 注意
    - *anti-examples* は programming/object-oriented design の分野で使われる用語とは無関係 (“反物質” (anti-matter) とのアナロジーで得た私の造語)

# “実例” だけで十分か?

- 固定された資料を使うことの根本的な問題
  - “正例の不足” の問題
    - 資料に可能性のすべてが実現されているわけではない
  - “負例の欠落” の問題
    - 資料には “正例” しか存在しない
    - どんなに容認度が低くても資料にある実例はすべて “正例”

# “実例” だけで十分か?

- (理論)言語学にとっては “負例の欠落” も問題
  - “正例の不足” は資料の規模が大きるとそれに応じて改善されるが, “負例の欠落” の解消率は解消されない
- 蛇足
  - 一部のコーパス言語学者が言うように「経験科学としての言語学に作例は不用である」ならば, それは逆説的に言語学の可能性を限定し, 最悪の場合には形骸化させる (少なくともそれは言語の認知科学にはならない)

# 一般化能力の自明視は危険

- ヒトは見かけは正例のみから言語を習得するが，習得の際に負例が必要でないことは必ずしも含意されない
  - 正例のみからの頻度主義の学習では説明しにくい事実がある
    - Fast Mapping (日高 & Smith 07, 08) や低頻度語の効率的習得
- 一般に “素朴” な機械学習はヒトの一般化をうまくシミュレートしない
  - 過剰般化をうまく回避する仕組み (e.g., 強化学習 (Sutton & Barto 98), Memory-based Learning (Daelemans & van den Bosch 05); 帰納推論 (坂本 & 中川 08)) が備わっていると考えないと認知発達の事実の多くはうまく説明できない

# 実例と作例の使い分け案

- 実例の収集を内観ベースで行なうのは非効率的 (かつ危険)
  - 十分に代表性のあるコーパスと，すぐれた検索ツールがあるなら，それを使った方がずっと早く，ずっと信頼性のある結果が得られる
- 作例は作例を使わないとできないこと (e.g., 反実例を使った議論, 統制された実験刺激の作成) に限定すると良い
- Word Sketch Engine (Kilgarriff & Tugwell 01, Srdanovic, *et al.* 08, スルダノビッチ・仁科 08) を使った実践例を以下で示す

# Sketch Engine を使った実例調査

# (Word) Sketch Engine の利点

- 言語学者が知りたいと思っている情報 (e.g., 文法関係の観点から見た共起情報) が，SkE を使うと効率的に抽出可能
  - 言語処理で開発された諸ツールと違って，始めから辞書編纂の目的に合うように開発された経緯をもつ
- SkE で利用できる JpWaC は大規模で実例に代表性を期待できる
  - 409,384,405 形態素 (BNC の約4倍) の Web コーパス
  - 詳細は http://nl.ijs.si/et/talks/CoJaS–7/lkaho.ppt

| | | |
|---|---|---|
| 歯 | 284 | 9.63 |
| 感性 | 115 | 8.67 |
| センス | 70 | 8.09 |
| 腕 | 139 | 8.04 |
| スキル | 73 | 7.73 |
| 原石 | 17 | 7.35 |
| 技 | 64 | 7.17 |
| 内面 | 25 | 6.82 |
| テクニック | 27 | 6.66 |
| 靴 | 46 | 6.61 |
| 己 | 17 | 6.55 |
| 技量 | 12 | 6.54 |
| 芸 | 17 | 6.54 |
| 技能 | 21 | 6.39 |
| 魂 | 29 | 6.04 |
| グラス | 11 | 5.95 |
| 感覚 | 58 | 5.92 |
| 歯磨き粉 | 6 | 5.92 |
| 教養 | 13 | 5.89 |
| 知性 | 11 | 5.83 |

| | | |
|---|---|---|
| ゴシゴシ | 6 | 9.31 |
| ごしごし | 5 | 9.17 |
| 更に | 9 | 6.97 |
| さらに | 30 | 6.83 |
| ひたすら | 5 | 6.38 |
| 常に | 8 | 5.95 |
| もう少し | 5 | 5.28 |
| もっと | 13 | 5.11 |
| ちゃんと | 13 | 5.04 |
| しっかり | 12 | 4.43 |
| いつも | 6 | 4.23 |
| きちんと | 5 | 3.77 |
| よく | 9 | 3.6 |

| | | |
|---|---|---|
| 歯ブラシ | 12 | 8.47 |
| すり | 8 | 8.22 |
| ワックス | 5 | 7.37 |
| ブラシ | 5 | 6.9 |
| 布 | 7 | 5.55 |
| ド | 5 | 2.64 |
| 環境 | 5 | 1.15 |
| 中 | 42 | 1.01 |
| 仕事 | 9 | 0.75 |
| 自分 | 11 | 0.1 |
| ここ | 5 | 0.07 |

| | | |
|---|---|---|
| 磨く | 55 | 7.37 |
| 光る | 32 | 6.81 |
| 上げる | 251 | 6.78 |
| 磨ける | 5 | 5.99 |
| 光らせる | 6 | 5.86 |
| 洗う | 23 | 5.68 |
| 輝く | 23 | 5.6 |
| 合える | 5 | 5.26 |
| 怠る | 7 | 4.97 |
| 抜く | 18 | 4.73 |
| 高める | 22 | 4.57 |
| 鍛える | 6 | 4.41 |
| たてる | 6 | 4.4 |
| かかる | 60 | 4.17 |
| 役に立つ | 6 | 3.43 |
| 積む | 5 | 3.35 |
| 役立つ | 7 | 3.33 |
| 育てる | 10 | 3.22 |
| 努める | 5 | 2.93 |
| たる | 5 | 2.91 |

| | | |
|---|---|---|
| ぬく | 21 | 7.54 |
| こむ | 17 | 5.22 |
| なおす | 6 | 4.31 |
| ゆく | 12 | 3.59 |
| いく | 222 | 3.36 |
| なさる | 14 | 3.33 |
| 続ける | 48 | 3.29 |
| 合う | 15 | 3.07 |
| 抜く | 5 | 2.95 |
| あう | 5 | 2.87 |
| いける | 29 | 2.79 |
| だす | 6 | 2.68 |
| おく | 37 | 2.64 |
| あげる | 13 | 2.33 |
| くださる | 30 | 2.2 |
| もらう | 25 | 2.08 |
| くれる | 33 | 1.71 |
| くる | 84 | 1.63 |
| 下さる | 8 | 1.55 |
| てる | 54 | 1.52 |

---

| nounは | 349 | 2.3 |
|---|---|---|
| スキル | 6 | 4.53 |
| 表面 | 6 | 4.13 |
| ひと | 5 | 3.66 |
| あと | 7 | 3.02 |

| nounに | 366 | 1.1 |
|---|---|---|
| 一生懸命 | 5 | 6.3 |
| 感性 | 6 | 4.92 |
| 為 | 7 | 3.14 |
| 一緒 | 10 | 2.79 |

| nounが | 247 | 1.0 |
|---|---|---|
| センス | 13 | 6.27 |
| 感性 | 12 | 5.95 |
| 能力 | 7 | 2.13 |
| 自身 | 7 | 1.94 |

# “磨く”のWord Sketch

# 掃除 JpWaC freq = 9888

change options

| prefix | 1945 | 14.6 |
|---|---|---|
| 大 | 1064 | 7.95 |
| お | 863 | 6.12 |
| 小 | 5 | 2.3 |

| をverb | 2504 | 3.4 |
|---|---|---|
| 済ませる | 22 | 6.09 |
| さぼる | 8 | 6.01 |
| すませる | 7 | 5.71 |
| サボる | 7 | 5.58 |
| 手伝う | 13 | 5.26 |
| 終える | 18 | 4.53 |
| 怠る | 5 | 4.45 |
| 頼む | 11 | 3.79 |
| こなす | 5 | 3.61 |
| 始める | 48 | 3.45 |
| はじめる | 6 | 2.82 |
| やる | 60 | 1.9 |
| する | 1250 | 1.77 |
| くれる | 33 | 1.71 |
| もらう | 19 | 1.67 |

| pronomの | 2004 | 2.7 |
|---|---|---|
| 魔法 | 109 | 8.74 |
| 年末 | 56 | 8.2 |
| 毎日 | 116 | 8.04 |
| 部屋 | 261 | 7.45 |
| 換気扇 | 9 | 6.94 |
| お部屋 | 14 | 6.87 |
| 冷蔵庫 | 19 | 6.73 |
| 兵舎 | 7 | 6.68 |
| 天下 | 10 | 6.63 |
| 網戸 | 6 | 6.43 |
| トイレ | 31 | 6.4 |
| 墓 | 23 | 6.32 |
| 便所 | 6 | 6.1 |
| 小屋 | 11 | 6.07 |
| 落ち葉 | 5 | 6.0 |

| がAdj | 176 | 2.1 |
|---|---|---|
| 楽 | 12 | 5.13 |
| 大変 | 33 | 4.77 |
| 面倒 | 9 | 4.7 |
| 嫌い | 8 | 4.48 |
| 簡単 | 10 | 3.5 |
| 大好き | 5 | 3.44 |
| 楽しい | 6 | 2.24 |
| やすい | 11 | 1.81 |
| 好き | 7 | 1.54 |
| 必要 | 18 | 1.33 |

| suffix | 1547 | 2.0 |
|---|---|---|
| 機 | 1101 | 9.14 |
| 術 | 112 | 7.93 |
| 婦 | 35 | 7.46 |
| 係 | 11 | 5.53 |
| どころ | 5 | 4.26 |
| 嫌い | 7 | 4.13 |
| 隊 | 8 | 3.85 |
| 用 | 36 | 3.56 |
| 編 | 8 | 3.33 |
| 好き | 19 | 2.96 |
| 器 | 6 | 2.85 |
| 等 | 28 | 1.63 |
| 中 | 50 | 1.26 |
| 後 | 12 | 1.06 |
| 法 | 9 | 0.48 |

| particle | 649 | 1.6 |
|---|---|---|
| なんか | 9 | 4.85 |
| なんて | 14 | 3.97 |
| だって | 8 | 3.75 |
| など | 108 | 3.22 |
| って | 19 | 2.79 |
| ばかり | 7 | 2.67 |
| も | 336 | 2.61 |
| だけ | 20 | 1.53 |
| と | 28 | 1.05 |
| という | 23 | 0.94 |

| はAdj | 81 | 1.4 |
|---|---|---|
| 苦手 | 9 | 4.88 |
| 大変 | 6 | 2.32 |

| はverb | 210 | 1.1 |
|---|---|---|
| 終わる | 13 | 1.94 |
| やる | 22 | 0.46 |
| 出来る | 6 | 0.08 |

| modifier Adj | 70 | 2.0 |
|---|---|---|
| 軽い | 11 | 4.71 |

| のpronom | 727 | 1.0 |
|---|---|---|
| コツ | 115 | 9.78 |
| おばさん | 37 | 8.19 |
| おば | 21 | 7.25 |
| オバサン | 5 | 7.15 |
| 手伝い | 9 | 6.37 |
| 仕上げ | 5 | 5.72 |
| おじさん | 10 | 5.67 |
| 最中 | 5 | 5.47 |
| 合間 | 6 | 5.28 |
| 仕方 | 31 | 5.26 |

| がverb | 278 | 0.9 |
|---|---|---|
| 行き届く | 26 | 8.6 |
| 済む | 7 | 3.88 |
| 終わる | 37 | 3.45 |
| 始まる | 10 | 2.19 |
| 出来る | 21 | 1.88 |
| できる | 38 | 0.26 |

| modifier Adv | 48 | 0.8 |
|---|---|---|
| 簡単 | 6 | 2.47 |

| にverb | 243 | 0.6 |
|---|---|---|
| 取り掛かる | 6 | 7.06 |
| 取りかかる | 5 | 6.64 |
| 追う | 6 | 2.81 |
| かかる | 9 | 1.46 |
| 始まる | 6 | 1.45 |
| 来る | 16 | 0.92 |
| 行く | 14 | 0.23 |
| 使う | 12 | 0.21 |

| でverb | 120 | 0.6 |
|---|---|---|

“掃除”のWord Sketch

# 現状の Sketch Engine の難点

1. サ変名詞は動詞用法が取り出せない
2. 語彙素/形態素認識の精度はまだ (英語に較べたらまだまだ) 十分ではない
3. 複合動詞のコロケーションが発見できない
4. 同音異義語が区別されていない
5. 異表記が統一されていない
6. (残念ながら無料ではない!!)

# 現状の Sketch Engine の難点 1/5

- サ変名詞は動詞用法が取り出せない
  - 名詞としての“結婚”の用法は抽出できるが，“結婚する”の動詞用法が抽出されていない
    - これはかなり痛いので，今後の改善に期待したい
  - Word Sketch の [ pronounの ]* で代用するしかないが，相当のノイズが混入する
    - *なぜ pronoun なのかは不明 (おそらくバグ)

# 現状の Sketch Engine の難点 2/5

- 語彙素/形態素認識の精度はまだ (英語に較べたらまだまだ) 十分ではない
  - 明らかな解析誤りは5%から25%ほど(差は品詞によるが) 存在する．例えば
    - “クローゼット” => “ゼット”
  - (形態素解析プログラム ChaSen 経由で) 日本語の記述文法が記述的に十分に妥当でない点が如実に表われている
    - 日本語で Coord(ination) が弱い理由は，複合動詞の扱いの不統一性に由来

# 現状の Sketch Engine の難点3/5

- 現状では
  - “ぬぐい+去る” < “ぬぐい+取る” << “?*ぬぐい+落とす”
  - “消し+去る” < “?消し+取る” << “??消し+落とす”
- の差を生む複合動詞のコロケーションが発見困難
  - 複合動詞は解析プログラムの段階で全部を前項と後項に分離して解析してくれないと有意義な一般化は不可能だが，解析プログラムを開発するNLP研究者はこれは複雑性を増やすだけなので，やりたくない
  - この点は言語学者が積極的に介入しない限り，絶対に改善されない

# 現状のSketch Engineの難点 4, 5/5

- 同音異義語が区別されていない
  - 用例分類のノイズになりやすい
- 異表記が統一されていない
  - 異表記の多い語彙素は，実例が分散するので相対的にサンプリング不足を招きやすく，結果的に精度低下がもたらされやすい

# 調査法と結果

# 調査の目的

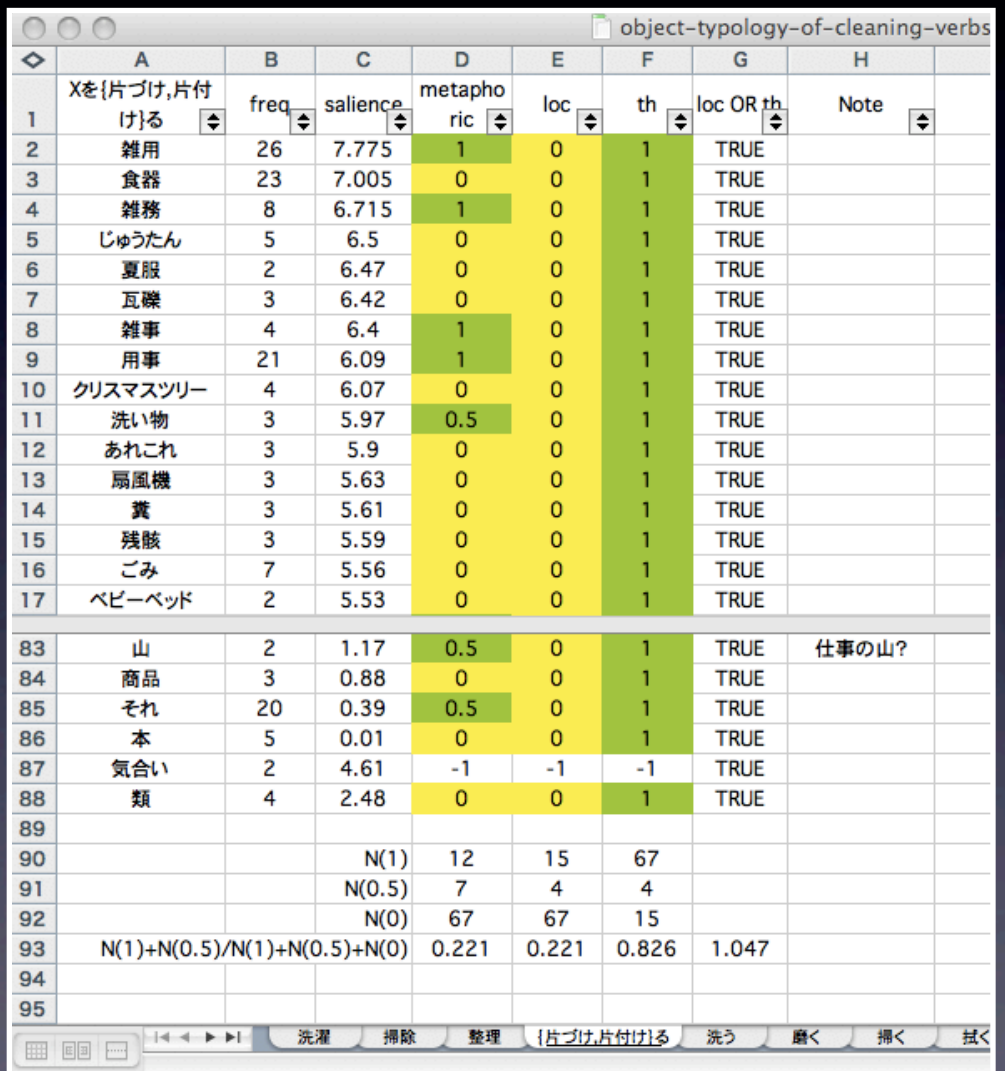
- 大谷の研究成果が日本語にどれほど転用できるかを見るため，次の9個の日本語の清掃動詞/サ変名詞 (*V*) のヲ格の名詞 (*X*) の意味的タイポロジーを調べる
  - “洗う”,“拭く”,“掃く”,“磨く”,“片{付け,づけ}る”,“落とす”
  - “掃除”,“洗濯”,“整理”
- 特に
  - メタファー用法のThバイアス源としての英語の不変化詞 (e.g.,*off, away, up*) に対応するものがあるか
- どうかに興味

# 調査の方法

- 名詞 *X* と動詞/サ変名詞 *V* の共起の強い組合わせを Word Sketch で収集
  - 動詞では [nounを], 名詞では [pronounの]* を見る (収集した用例の数は 10 から 50)
- *V* とヲ格名詞 *X* の組み (*V*, *X*) の人手コーディング
  - Metaphoric = {1, 0.5, 0}: 事例がメタファーかどうか
  - Th = {1, 0}: *X*が<除去する対象>を表わすかどうか
  - Loc = {1, 0}: *X*が<除去の対象>の存在する<場所>あるいはモノを表わすかどうか

# コーディングの見本

- “片付ける” と “片づける” の用例を統合したもの
- freq, salience は Sketch Engine の出力をそのまま記載
- 結果は次の URL から入手可能:
  - http://clsl.hi.h.kyoto-u.ac.jp/~kkuroda/data/object-typology-of-cleaning-verbs.xls



object-typology-of-cleaning-verbs

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | Xを{片づけ,片付け}る | freq | salience | metaphoric | loc | th | loc OR th | Note |
| 2 | 雑用 | 26 | 7.775 | 1 | 0 | 1 | TRUE | |
| 3 | 食器 | 23 | 7.005 | 0 | 0 | 1 | TRUE | |
| 4 | 雑務 | 8 | 6.715 | 1 | 0 | 1 | TRUE | |
| 5 | じゅうたん | 5 | 6.5 | 0 | 0 | 1 | TRUE | |
| 6 | 夏服 | 2 | 6.47 | 0 | 0 | 1 | TRUE | |
| 7 | 瓦礫 | 3 | 6.42 | 0 | 0 | 1 | TRUE | |
| 8 | 雑事 | 4 | 6.4 | 1 | 0 | 1 | TRUE | |
| 9 | 用事 | 21 | 6.09 | 1 | 0 | 1 | TRUE | |
| 10 | クリスマスツリー | 4 | 6.07 | 0 | 0 | 1 | TRUE | |
| 11 | 洗い物 | 3 | 5.97 | 0.5 | 0 | 1 | TRUE | |
| 12 | あれこれ | 3 | 5.9 | 0 | 0 | 1 | TRUE | |
| 13 | 扇風機 | 3 | 5.63 | 0 | 0 | 1 | TRUE | |
| 14 | 糞 | 3 | 5.61 | 0 | 0 | 1 | TRUE | |
| 15 | 残骸 | 3 | 5.59 | 0 | 0 | 1 | TRUE | |
| 16 | ごみ | 7 | 5.56 | 0 | 0 | 1 | TRUE | |
| 17 | ベビーベッド | 2 | 5.53 | 0 | 0 | 1 | TRUE | |
| 83 | 山 | 2 | 1.17 | 0.5 | 0 | 1 | TRUE | 仕事の山? |
| 84 | 商品 | 3 | 0.88 | 0 | 0 | 1 | TRUE | |
| 85 | それ | 20 | 0.39 | 0.5 | 0 | 1 | TRUE | |
| 86 | 本 | 5 | 0.01 | 0 | 0 | 1 | TRUE | |
| 87 | 気合い | 2 | 4.61 | -1 | -1 | -1 | TRUE | |
| 88 | 類 | 4 | 2.48 | 0 | 0 | 1 | TRUE | |
| 89 | | | | | | | | |
| 90 | | | N(1) | 12 | 15 | 67 | | |
| 91 | | | N(0.5) | 7 | 4 | 4 | | |
| 92 | | | N(0) | 67 | 67 | 15 | | |
| 93 | N(1)+N(0.5)/N(1)+N(0.5)+N(0) | | | 0.221 | 0.221 | 0.826 | 1.047 | |
| 94 | | | | | | | | |
| 95 | | | | | | | | |

洗濯 | 掃除 | 整理 | {片づけ,片付け}る | 洗う | 磨く | 掃く | 拭く

# 蛇足

- ゴリゴリのコーパス言語学者や言語処理関係者には「人手コーディングは主観性が混入するからダメだ」と言う人がいるけど，これは本末転倒
  - PoS tagging や (tree) parses が完全に客観的だと思うのはかなり深刻な勘違い
- 理由
  - 客観性と信頼性 and/or 代表性は本質的に別の指標であり，かつ，より重要なのは後者の方
  - 過度の客観性の要求は萌芽な段階にある経験科学の発展を阻害する

# 調査から得られた一般化

A. メタファー表現は慣用化/定型化する傾向がある
B. 結果を含意する動詞がメタファー用法をもちやすい?
C. Th 選好の強い動詞でメタファー用法が定着しやすい?
D. 英語の不変化詞の一つ (e.g, *off*) に，日本語では複数の複合動詞の後項が対応 (e.g,“取る”,“落とす”,“去る”)し，面白い並行性もある
E. Loc/Th は排他的とは限らない?

# 一般化 A の評価

A. メタファー表現は慣用化/定型化する傾向がある

- 証拠: Salience = (Mutual Information * Log Frequency) の高い語句の組合わせを表現した Word Sketch の結果にメタファーの例が多いということ自体から示唆
  - *X*を磨く:
    - *X* ⇒ {腕, 技能, 芸, 技, スキル, ...}, {心, 内面, 自分, 己, ...}
  - *X*を {片付け,片づけ}る:
    - *X* ⇒ {用事, 用件} ⇒ {仕事, タスク, 家事, 雑用, 問題, 案件};
    - 雑用 ⇒ {雑務, 雑事}

# 一般化 B の評価

- Loc か Th の値が+である X の個数を N
  - M指数 = Metaphoric の値が 1 か 0.5 の X の個数 / N
  - “洗濯”のM指数は“命の洗濯”に限られる (e.g., ?*命を洗濯しながら)
- <結果の含意>の有無は直観に基づいてコーディングしているが, “V1てV2” の分布指標を用いて数値化も可能なハズ
- <和語に限り>という限定をつけても “洗う” が例外となる

| V | M指数 | 結果の含意 | 支持 |
|---|---|---|---|
| 磨く | 0.793 | ++ | Yes |
| 落とす | 0.733 | + | Yes |
| 整理 | 0.483 | ++ | Yes |
| 片{付け,づけ}る | 0.224 | ++ | Yes |
| 洗濯 | 0.272* | + | Yes* |
| 洗う | 0.02 | + | No |
| 掃除 | 0.02 | ++ | No? |
| 拭く | 0 | – | Yes |
| 掃く | 0 | – | Yes |

# 一般化 C の評価

- Loc か Th の値が 1 か 0.5 である X の個数を N
  - Loc 指向指数 = Loc の値が 1 か 0.5 の X の個数 / N
  - Th 指向指数 =Th の値が1か 0.5 の X の個数 / N
- “整理” のメタファー用法の場合，Th と Loc の区別は非排他的

| 動詞 | Loc指向指数 | Th指向指数 | M指数 | 支持 |
|---|---|---|---|---|
| 磨く | 1.0 | 0 | 0.793 | No |
| 落とす | 0 | 1.0 | 0.733 | Yes |
| 整理 | 0.501 | 0.649 | 0.508 | Yes |
| 洗濯 | 1.0 | 0 | 0.25* | No? |
| 片付ける | 0.221 | 0.826 | 0.221 | Yes |
| 洗う | 0.978 | 0.022 | 0.022 | Yes |
| 掃除 | 0.96 | 0.04 | 0.02 | Yes |
| 掃く | 0.833 | 0.167 | 0 | Yes |
| 拭く | 0.783 | 0.217 | 0 | Yes |

# 一般化 D の評価

| | -落とす | -去る | -取る | -飛ばす | -上げる | 結果指向性 |
|---|---|---|---|---|---|---|
| たたき,叩き | ++ | – | 0 | 0 | 0* | – |
| はたき | ++ | – | – | + | – | – |
| 洗い | ++ | – | 0 | + | 0 | + |
| ふき, 拭き | + | – | ++ | – | 0 | + |
| こすり | ++ | – | ++ | – | – | – |
| ぬぐい,拭い | 0 | + | ++ | – | – | + |
| はき, 掃き | – | 0 | 0 | + | 0 | + |
| 消し | – | + | 0 | 0 | – | + |
| 磨き | – | – | – | – | ++ | + |
| 片{づけ,付け} | – | 0 | – | – | 0 | + |
| (だまし,騙し) | – | – | ++ | – | – | + |

- 内観ベースの結果
  - + は複合動詞 (e.g., 叩き落とす) が存在すること，– はしないこと．0 はあってもおかしくないを表わす
  - {++, +, 0, –} はコーパス頻度でより正確に推定可能
- “落とす”,“飛ばす”,“去る”,“取る” が *off* や *away* に対応?
- “上げる” が *up* に対応
- Loc 選好 vs Th 選好の傾向は日英語で共通
- 後項動詞がメタファーを認可する傾向も共通

# 不変化詞と複合動詞の対応

- Talmy のタイポロジー (Talmy 75, 76, 85, 03) で英語は Satellite-frame L で，日本語は Verb-frame Lなので
- VfL では不変化詞は複合動詞で表わされ，
- その組合わせには (コロケーションに由来し，意味論的に説明できるとは限らない) 面倒な制限がある
- 後者の可能な説明
  - SfLで不変化詞は10個のオーダーでしか存在しないのに対し，VfLで不変化詞に相当する動詞は100個のオーダーで存在し(しかも文法化や用法の衰退によって数が変動する)
  - SfL では経路を表わすのが不変化詞であるのに対し，VfL では経路を表わすのが動詞 (Matsumoto 03)

# Vの意味とXの役割の共選択

- 清掃動詞 *V* (e.g.,“洗う”) の目的語名詞 *X* が <th(ing to be removed)>か <th の付着する loc(ation)> かが問題

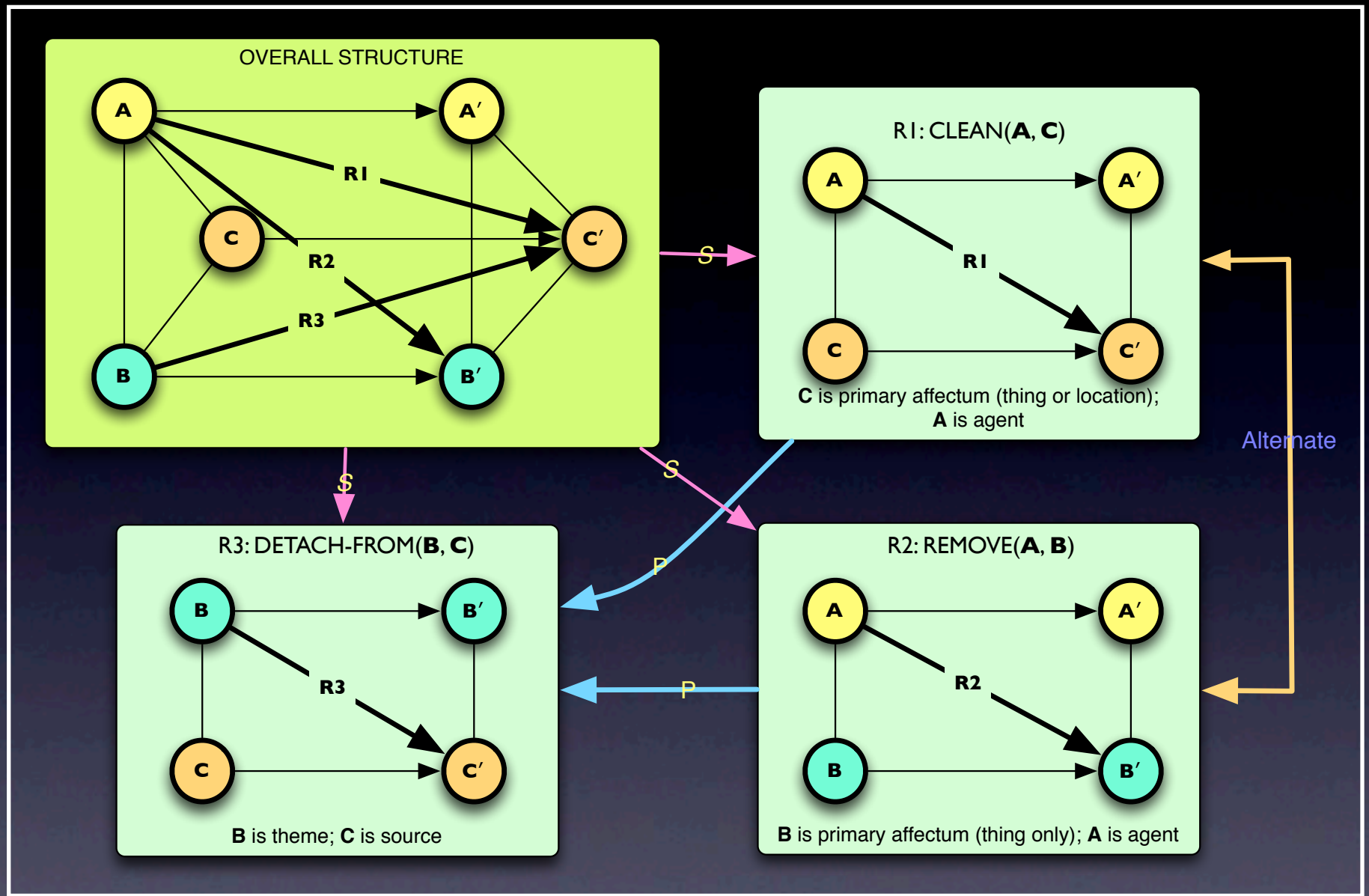
(1) (<loc:服>の) <th:汚れ>を 洗う

(2) <loc:服>を 洗う [cf. *<loc:服>から <th:汚れ>を 洗う]

(3) a. (<loc:服> {の; から}) <th:汚れ>を洗い落とす; b. *<loc:服>を洗い落とす

(4) a. (<loc:服>{の; から})<th:汚れ>を落とす; b. #<loc:服>を落とす

- 注意: th は意味役割の Theme と一致することもある (e.g., (1), (3a), (4a)) が, 常にというわけではない




ヲ格名詞 X は R1の C を表わす場合も R2の B を表わす場合もある

R1が primary/foreground の時, X = C, R2 が primary/foreground の時, X = B,

X は (B だろうと C だろうと) 常に primary affectum?

# 一般化 E の評価

E. Loc/Th は排他的とは限らない?

- 由来:“Xを磨く” のメタファー用法で X が Loc か Th かはっきりしないことから思いついた仮説
  - ただ “Xを磨く” の字義通りの用法で Xが<除去対象>を表わすものはない
  - 類似の現象:“Xを整理する” のメタファー用法で X が Loc か Th か判別不能
- 発展的疑問:“Xを磨く” で X は常に Loc だが, どうしてメタファー用法で X に<産物> (product) の意味が出るのか?
  - R3:A clean C が primary になることの副作用が理由の一つに考えられるが, この疑問は現時点で未解決

# 交替の条件と<産物>の含意

- X が Th とか Loc とか言うのは，次のフレームごとに個別に決まるこ
  - F1: <A remove B from C>
  - F2: <A clean C (of B> (is-a <A improve C>)
  - F3: <C detach-from C> (implied <B disappear-from C>)
    - {F1, F2, F3} は清掃表現の意味を構成する最低限のフレーム群
- F1 と F2 のどっちが primary になるかで揺れる現象が Loc/Th 交替
  - ヲ格として F1 の B が現われたり，F2 の C が現われたりする
- F3 で定義される Theme と Loc (= Theme の Source) は清掃の概念化に常在
- X が C=Th を表わすか B=Loc を表わすかに関係なくメタレベルで成立する意味役割が Affectum，これの Intended Result が Product

# 発表のまとめ

# “中道” を行くための方法論

1. 代表性のある資料 $R$ (e.g., コーパス) から，なるべく多くの正例集合 $P = \{p1, p2, ...\}$ を収集
2. $P$ を基に研究者が (直観 $I$ をうまく働かせて!!) 有意義な記述的一般化 $G(P) = \{g1, g2, ...\}$ を得る
   - 分布類似度に基づく自動分類の精度には限界があり，直観 $I$ は有意義な一般化のために不可欠
3. $G$ の妥当性を負例集合 $N = \{n1, n2, ...\}$を反実例として使って評価 (予測の確証あるいは反証)
   - この段階では作例が不可欠

# 補足的注意

- 現状の検索技術では効率的に抽出できない現象 (e.g, 介在性構文) があるのも確か
- そういう現象の実例は，実際に資料 (e.g., 小説) を読んで地道に探すしかない
  - コーパス利用は，そういう資料探索の特殊な場合
- 注意
  - 効率的に見つからない実例を得ようとして，作例で代用するのは危険
    - 確証バイアスのため，知らないうちに“自作自演”をする危険がある

# 検索の限界を克服するために

- 最大の問題
  - 現行では form を key にした検索しかできない
    - parsed corpus の利用もこの範囲内
  - ⇒ sense を key にした検索ができない
  - ⇒ X構文の多くが検索できない
    - 例外は way 構文のように特定の語彙項目を含むもの
- 構文レベルの高次の検索が可能となるためには sense-tagged コーパスが不可欠
- 更に言えば，その先に role-tagged コーパスがあれば，更に便利

# 謝辞


- 次の方々からの意見が有益でした
  - 加藤鉱三 (信州大学)
  - 黒宮公彦 (大阪学院大学)

# References [1]


- Daelemans, W. and van den Bosch, A. (2005). *Memory-Based Language Processing*. Cambridge University Press.
- Kilgarriff, A. and D. Tugwell (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. Information Technology Research Institute Technical Report ITRI-01-12.
- Matsumoto, Y. (2003). Typologies of lexicalization patterns and event integration: Clarifications and reformulations. In S. Chiba, *et al.* (eds.), *Empirical and Theoretical Investigations into Language: A Festschrift for Masaru Kajita*, (pp. 403-418), Tokyo: Kaitakusha.
- Newmeyer, F. J. (2003). Grammar is Grammar and Usage is Usage. *Language* 79 (4): 682-707.

# References [2]


- Srdanovic Erjavec, I and Erjavec, T. and Kilgarrif, A. (2008). A web corpus and word sketches for Japanese. *J. of Natural Language Processing* 15/2.
- Sutton, R. S. and Barto, A. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.
- Talmy, L. (1975). Semantics and the syntax of motion. In J. Kimball (ed.), *Syntax and Semantics 4* (pp. 181-238), Academic Press.
- Talmy, L. (1976). Semantic causative types. In M. Shibitani (ed.), *Syntax and Semantics 6: The Grammar of Causative Constructions*. Academic Press, N.Y., pp. 43-116.
- Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (ed.), *Language Typology and Syntactic Description III: Grammatical Categories and the Lexicon* (pp. 57-149), Academic Press.
- Talmy, L. (1991). Path to realization. *BLS* 17, 480-519.

# References [3]


- 日高 昇平・Smith, L. B. (2008). 自然物体の “種類” に固有な新規後の汎用. 第25回日本認知科学会発表論文集.
- 坂本 佳陽・中川 正宣 (2008). 帰納推論の計算モデルが明らかにする人格と状況の相互作用. 第25回日本認知科学会発表論文.
- スルダノヴィッチ-エリャヴェッシ, I・仁科 喜久子 (2008). コーパス検索ツール Sketch Engine の日本語版とその利用方法. 日本語科学 24: 59-80.