意味フレーム階層ネットワークの効率的構築

ーシソーラスを用いた上位語への置き換えを利用してー

京都大学大学院
金丸 敏幸

日本認知言語学会　設立五周年記念全国大会

はじめに

- 目的
- FOCALの構築方法
- 自動化の理由
- 今回の手法の説明
- 考察

目的

- FOCAL構築の省力化のためのcase study。
- 有効性とその問題点を探る。

FOCALの構築作業

- 黒田[1]から抜粋
  1. 一行毎に主語句、目的語句を文脈から取り出す
  2. 主語句と目的語句の意味タイプを特定する。
  3. 主語句と目的語句の意味役割の特定
  4. 文の意味フレームの特定
  5. フレームのネットワーク構造の構成

FOCALの構築作業

- 黒田[1]から抜粋
  1. 一行毎に主語句、目的語句を文脈から取り出す
  2. 主語句と目的語句の意味タイプを特定する。
  3. 主語句と目的語句の意味役割の特定
  4. 文の意味フレームの特定
  5. フレームのネットワーク構造の構成

→ 非常にたいへん

自動化の理由

- 対象データの量が多い
- 記述の一貫性をねらう
- 記述の労力を減らす

対象データの量

- 一語あたりの対象データ数
- 分析の対象となる語の数

記述の一貫性

- これまでの言語学で欠けていたこと
  - 「網羅性」
  - 「一貫性」
- 自動化の際にフォーマットを統一。大量のデータを、同じ方法で記述できる。

記述と労力

- FOCALの現時点での問題点
  - 意味分析として有効な手段ではあるが、手間と時間がかかる。
  - 特に初期段階で見通しが立てにくい。
- これらを少しでも省力化することができれば、より多くの記述を行うことができる。

分析対象

- 以下のテキストデータから該当文をKWICで抽出。
  - Webテキスト:5月17日～20日にかけて採集。
  - メールマガジン:バックナンバー
  - 辞書テキスト
  - 新聞テキスト
  - 小説テキスト
- 合計　約 657 MB (約 3億4445万 字)

検索内容

- 検索内容　襲う
- /(が|に)襲(わ|い|う|え|っ)/を検索
- 検索該当数　1,103例

自動化方法

1. 格助詞の直前語に対し、重複語を削除。
2. 「日本語語彙大系」を用いて、直前語を一つ上位の語に置き換え。
3. 1. に戻る。

- 自動化作業には、プログラミング言語 Perl (http://aspn.activestate.com/ASPN/Downloads/ActivePerl/) を使用

日本語語彙大系

- 日本語の語彙を階層によって、分類
- 単語体系は，日本語の単語30万語に対して，各々の単語が持つ意味属性を示したものである. (日本語語彙大系 : 単語体系より)

「語彙大系」使用の理由

- カテゴリー体系が含有されている
  - 自動化作業に人間のカテゴリー化能力とその傾向を反映させることが可能
- 意味に基づいた情報圧縮が可能
  - 語彙が上位になるに従って、フレーム記述に近づくと考えられる
- 階層性を持っている
  - 情報圧縮をした後に、フレームの階層を記述することの省力化が可能

置き換えの条件

- 多義性の扱い
- 大系にない語の取り扱い

多義性の扱い

- 複数の意味属性に含まれる語については、そのどれにも置き換えを行う。
  - 後の作業で他の意味属性との統合を行うため、幅広く置き換えをしておいても、本来の意味からはずれた用法は、統合の段階で無視できるようになるはずという見込み

大系にない語の取り扱い

- 大系は約10万語の一般名詞を登録
  - しかし、これだけでは十分カバー出来ない
- 固有名詞について
  - 固有名詞辞書は使用しなかったため、固有名詞や複合名詞については、置き換えが行われず、分析の対象からはずれてしまっている

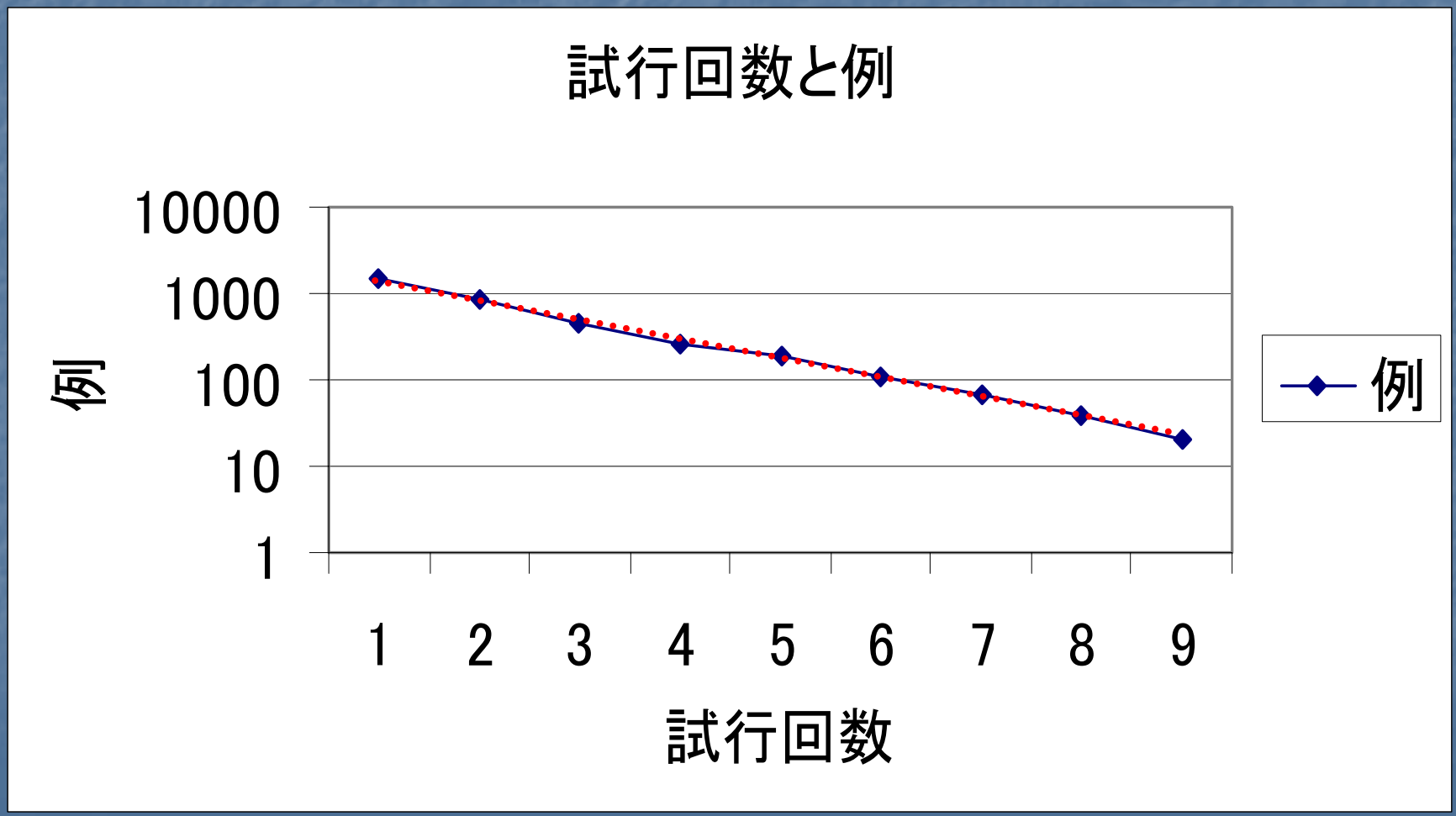
「襲う」による実験

- 最初の置き換えで、約1000例から約250例(重複削除済み)にまで、まとめられた。
- 出現している置き換え例
  - 病気類 風〔かぜ〕日照り･冷害等 難 弾効 音 土砂崩れ･雪崩等 風水害
  - 武器 形勢 速度 群 刃物 軍
  - 虫(その他) 病気類 爬虫類 形 魔物･化け物 獣
  - 感覚 気分 感動　恐れ 気配

「襲う」による実験

- 二回目の置き換えで、約250例から約190例(重複削除済み)にまで、まとめられた。
- 出現している置き換え例
  - 気象 天災 災難
  - 団体 軍・隊 動物(個体)
  - 焦躁・くつろぎ 同情・嫉妬
- このレベルになると主語句が基本レベルでカテゴリー化されている。

「襲う」による実験

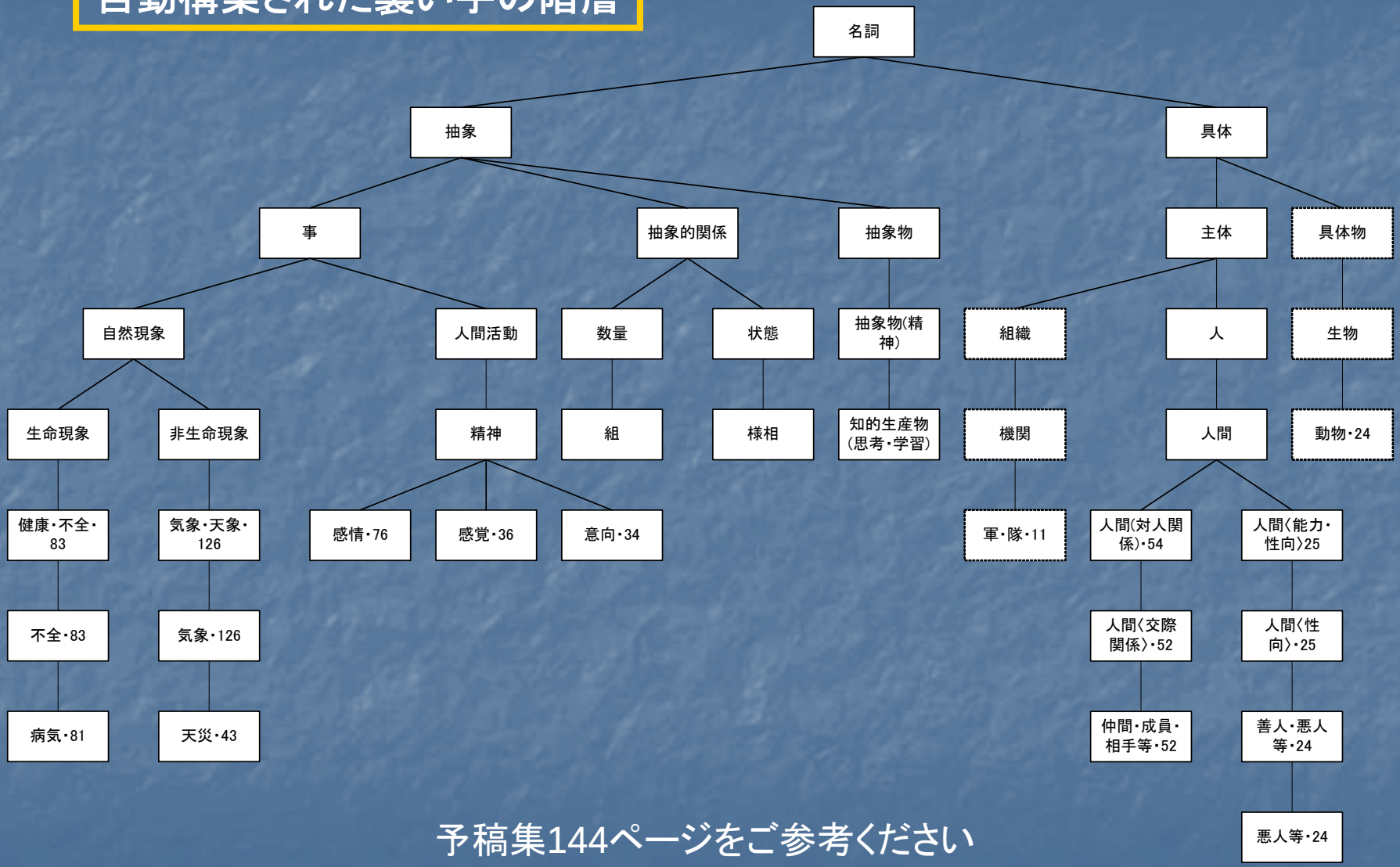


抽出されたフレームの例

- 今回抽出したのは、「襲い手」
  - 人間、仲間・成員・相手等、悪人等
  - 病気、天災、気象、意向、感覚、感情
  - 様相、組、知的生産物(思考・学習)

- どの語がフレームと関係するかは、語に依存する。

自動構築された襲い手の階層

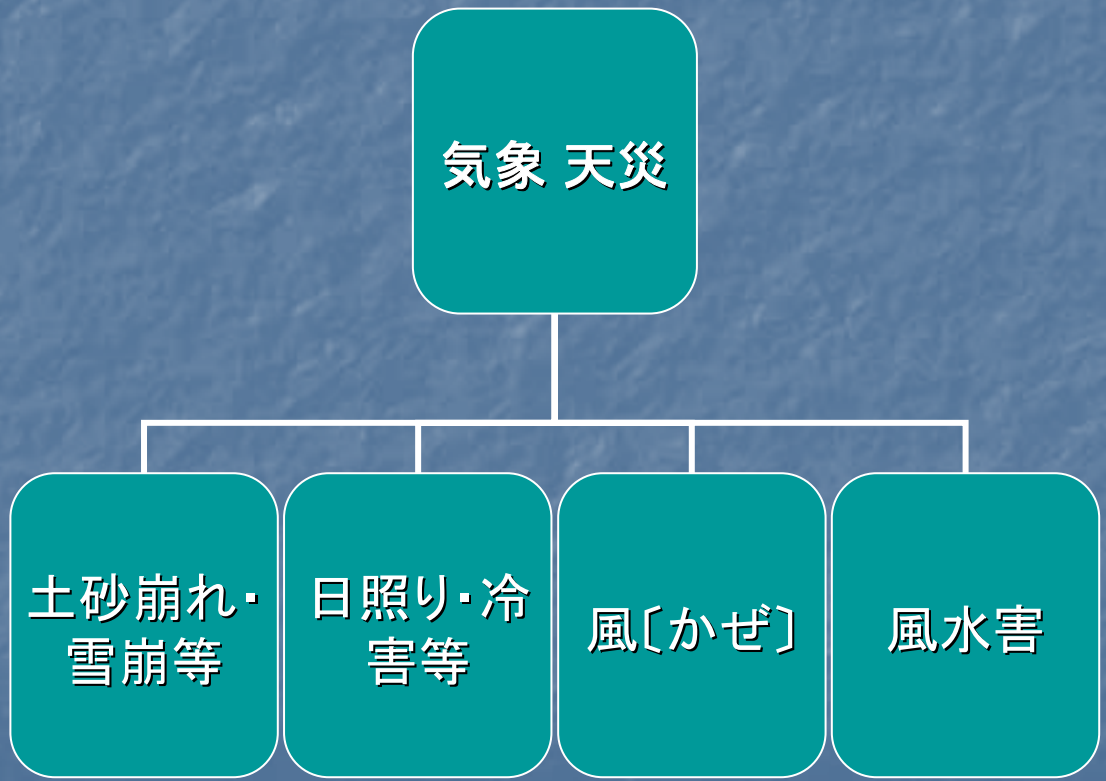


予稿集144ページをご参考ください

自動構築された襲い手の階層

予稿集144ページをご参考ください

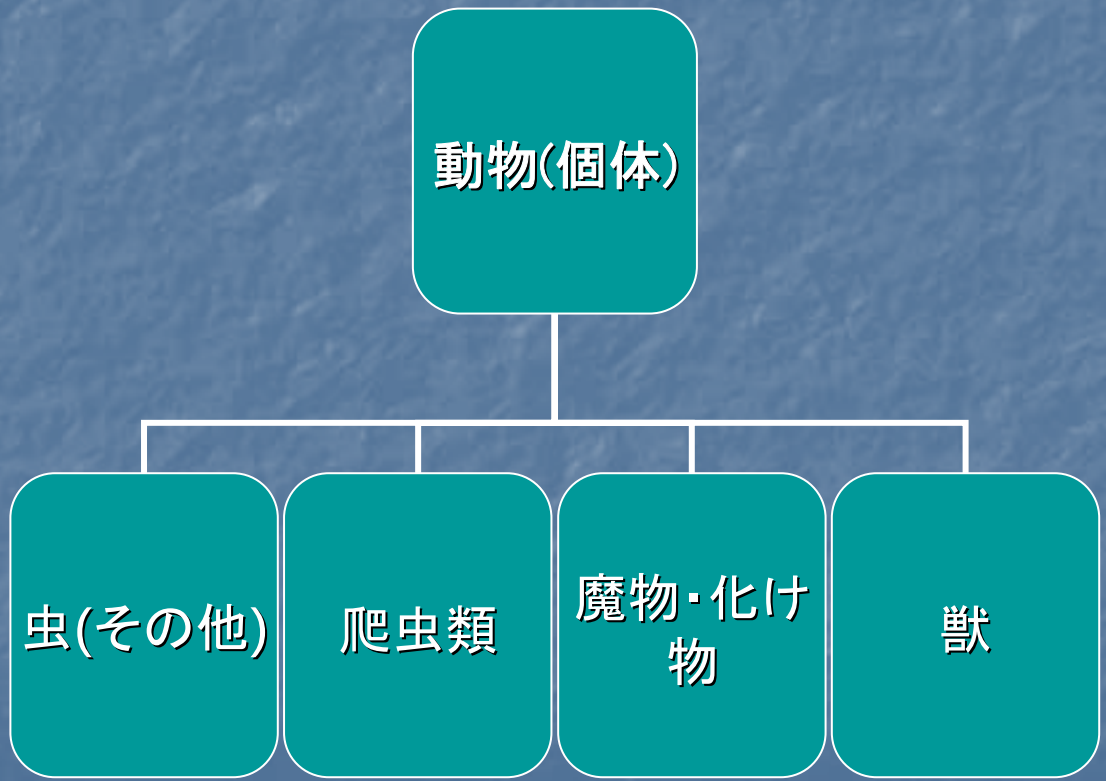
「襲う」による実験



←二回目に出現

←一回目に出現

「襲う」による実験



考察

- 大量のデータを扱える
  - 人手だけでは見落としがちなデータも発見可能
  - 今回の例では、「感情、感覚」が襲うというフレームを発見
- フレーム関係の特定
  - 「感情、感覚」という襲い手は、「病気」や「気象・天災」が襲うというフレームに近い物であることが予想可能

問題点

- 細かい粒度のフレーム分析
  - 動物への攻撃(捕食的)
  - 動物への攻撃(非捕食的)

- これら違いは、発見が困難
  - 同じカテゴリーに属する襲い手の区別は原理的に不可能

問題点

- 細かい粒度のフレーム分析　…NG.
  - 動物への攻撃(捕食的)
  - 動物への攻撃(非捕食的)

- 身体性を反映したフレーム　…OK.
  - 「手がかり」をつかむ → 探索、発見
  - 「足がかり」をつかむ → 前進、上昇

今後の課題

- 精度の向上
  - 自動的に得られた分類を基に、機械学習の手法を用いて、意味フレームの特定精度をあげる
- 多義語の扱い
  - 単語同士の類似度を利用したクラスター分析を用いて、有効でない分類を極力減らす

謝辞

- 独立行政法人 情報通信研究機構
  - 村田真樹氏
- 京都大学大学院山梨研究室のメンバ
- 本研究は、平成16年度科学研究費補助金(学術創成研究 課題番号: 13NP0301)の援助の下で行われた

対象データ

- http://amrita.s14.xrea.com/d/ 3.16 MB
- http://k-ryosha.jp/index.html 2.06 MB
- http://www.med-legend.com/ 1.38 MB
- http://www.isis.ne.jp/mnn/senya/ 89.2 MB
- http://tanakanews.com/ 5.21 MB
- http://www.asahi-net.or.jp/~FV6N-TNSK/ 4.31 MB
- http://alisato.parfait.ne.jp/diary/ 3.84 MB
- http://blue-brewery.net/ 316 KB
- http://www.finalbeta.jp/update_log/ 4.16 MB
- http://www.alpha-net.ne.jp/users2/ginjy/ 916 KB
- http://www-nishio.ise.eng.osaka-u.ac.jp/~ueda/diary/ 2.13 MB
- http://d.hatena.ne.jp/kanryo/ 1.45 MB
- http://neutron.tdiary.net/ 652 KB
- http://kiri.jblog.org/ 1.72 MB
- http://www.rubyist.net/~matz/ 2.69 MB
- http://homepage3.nifty.com/mogami/diary/ 447 KB
- http://www.pure.cc/~pramm/ 3.53 MB
- http://www.na.rim.or.jp/~achi-oya/hiroko/ 5.55 MB
- http://ww9.tiki.ne.jp/~cana/ 4.19 MB
- http://blog.tatsuru.com/ 4.76 MB
- http://zebra.s3.xrea.com/diary/ 2.00 MB

対象データ

- http://talk.to/noda/ 2.59 MB
- http://www.mori-office.com/ 8.82 MB
- http://www.moriyama.com/netscience/ 6.12 MB
- 世界大百科事典 マイペディア(日立システムアンドサービス) 148.87 MB
- 大辞林(三省堂) 123.48 MB
- スーパーニッポニカ2001(小学館) 116.15 MB
- 読売新聞･Dairy Yomiuri 対訳コーパス 66.8 MB
- 新潮文庫の百冊 CD-ROM版 41.1 MB

参考文献

- 黒田航、伊佐原均 2004「日本語の意味(役割) タグ体系を定義する試み -FrameNet の視点から」言語処理学会第10 回年次大会発表論文集: 148–152. 言語処理学会.
- 黒田航、中本敬子、野澤元 *未公開論文*「状況理解の単位としての意味フレームの実在性に関する研究」
- NTTコミュニケーション科学基礎研究所 1999「日本語語彙大系　CD-ROM版」岩波書店