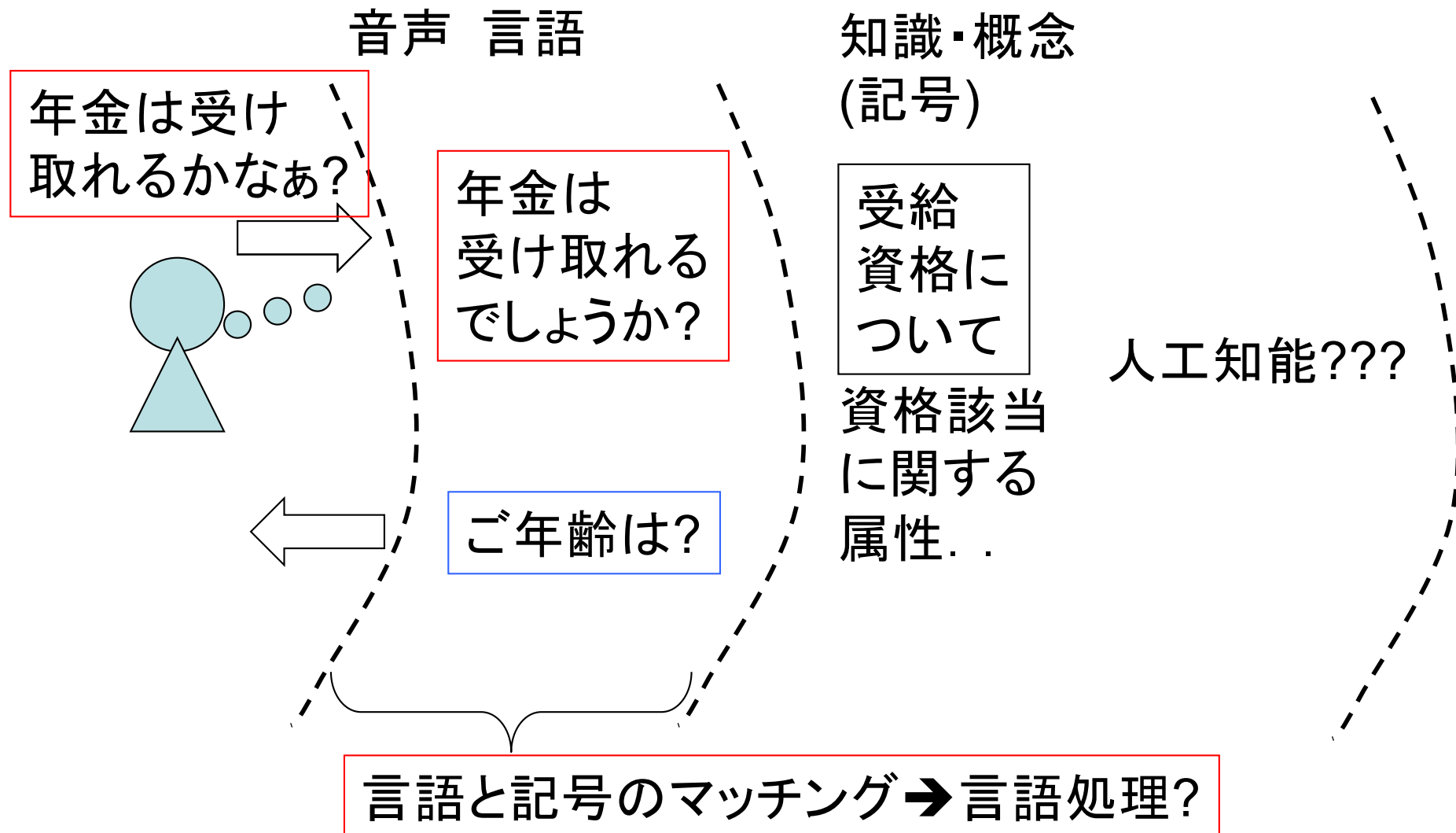


言語資源の今までと 言語処理研究のこれから

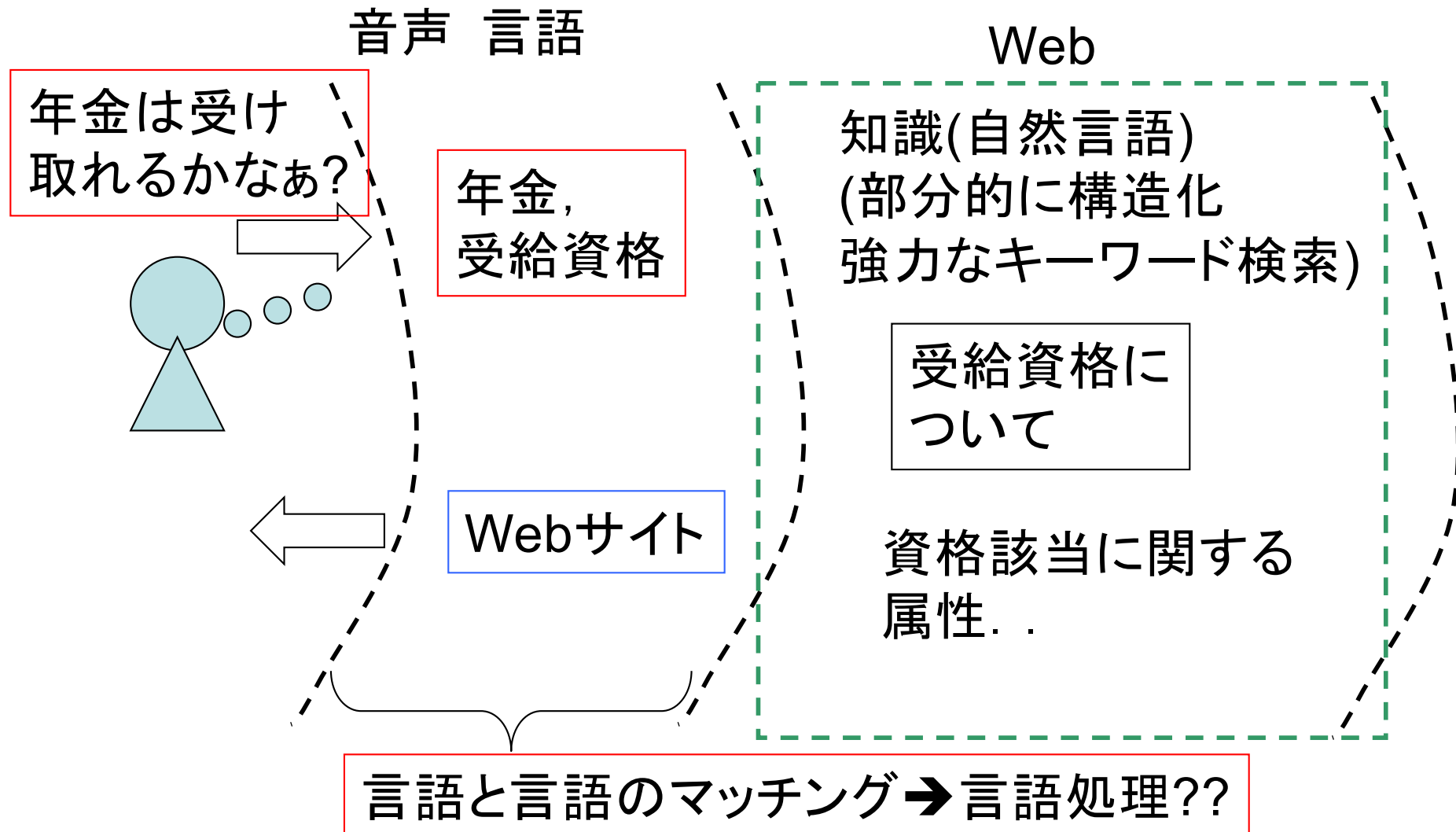
岡山大学 竹内孔一

音声言語シンポジウムの企画での
講演内容記録 2008年12月9日 早稲田大学にて

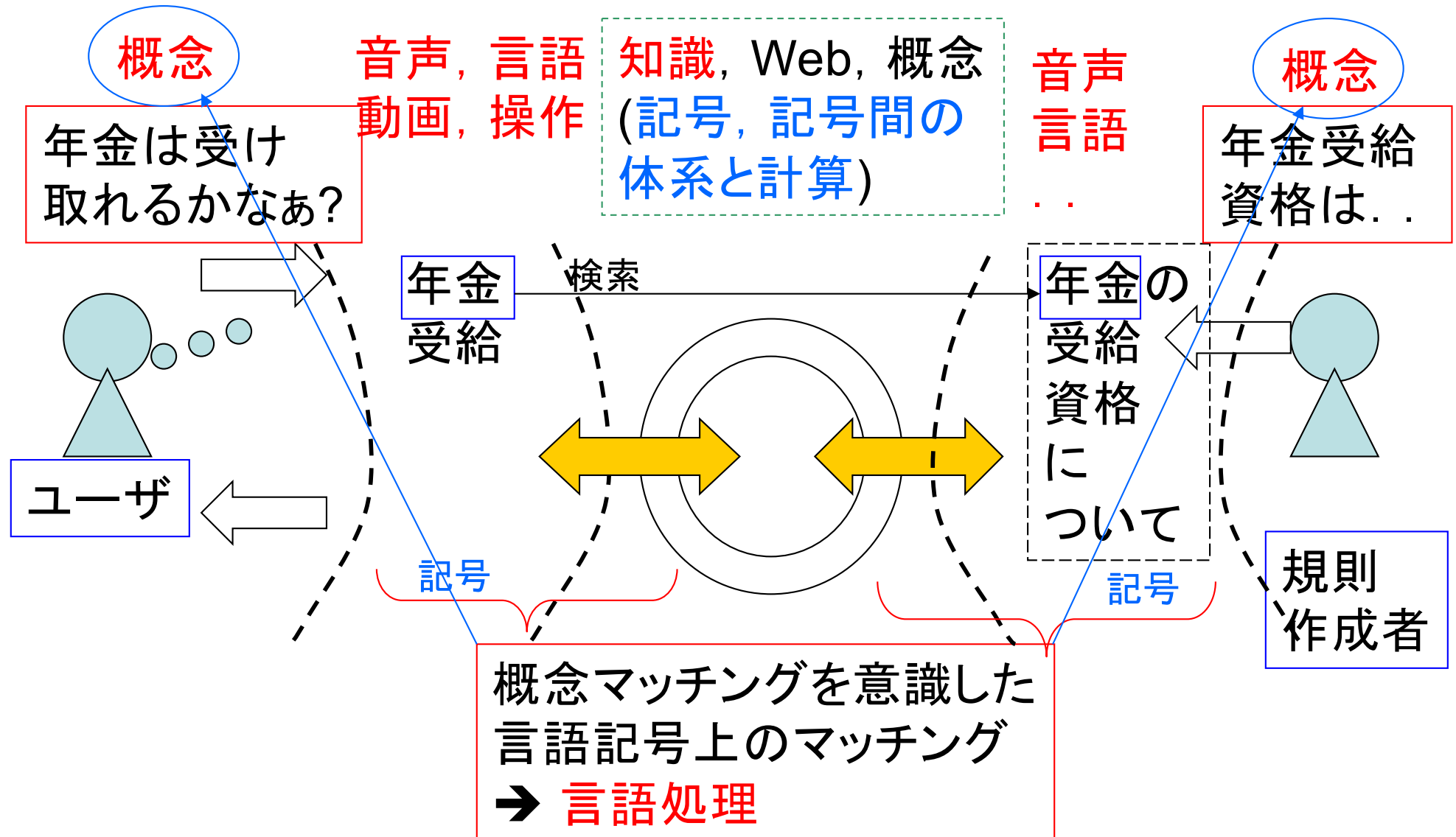
言語処理の位置づけ(1/3)



言語処理の位置づけ(2/3)



言語処理の位置づけ(3/3)



言語資源とは

• 言語資源とは？

- ある目的で収集された文書・語彙
 - テキストコーパス, 用語辞書
- ある定義に基づく意味タグが付与された文書・語彙
 - 語義付与コーパス, Genia コーパス
- 意味的關係づけがなされた文書・語彙
 - NTCIR の判定結果, WordNet, 対訳コーパス
- 定義そのもの
 - GDA (橋田97)

• 利用

- 統計・標本
 - 学習データ, 評価データ
 - 学習対象となる定義集合
- 定義集合
 - 辞書, 同義語集合
 - 処理目的の定義集合 (例) 語義集合

統計的学習モデル
と強い関係

言語資源の系列(1/3)

- 言語学・認知心理学から派生

言語理論の
具現化

- Levinの動詞交替辞書(93)
- LCS(95), VerbNet(06), FrameNet(98), WordNet(90)
- TimeML(04), MSFA(06), 動詞項構造シソーラス(08)

- 標本を意識(言語学心理学のため)

Webで配布

- Brown corpus (61) テキストデータの標本
- 日本語コーパス BCCWJ (08) 国語研

- 自然言語処理が中心

- Penn tree bank (92)(構文解析結果の付与)
- EDR (1995-): 日英中辞書タグ付与コーパス
- Propbank (02): 語義(意味役割) 付与テキスト
- SemEval(02): 日本語語義付与
- 京大コーパス(02), NAIST テキストコーパス(07)

言語資源の系列(2/3)

- タスク指向

- 分子生物学:
 - Genia コーパス(辻井研)
 - Gene ontology consortium: 同義語を整理
- 固有表現階層(関根), 鳥バンク(池原)
- TREC, NTCIR (NII) 検索データ, ACEコーパス

- Web2.0的

- wiki: 百科事典
- 青空文庫: テキストデータ
- Web上にあるテキストデータ, blog
- 英辞郎: 英日翻訳辞書
- Google n-gram データ

言語資源の系列?(3/3)

- タグの定義(XML)

- GDA (橋田97)

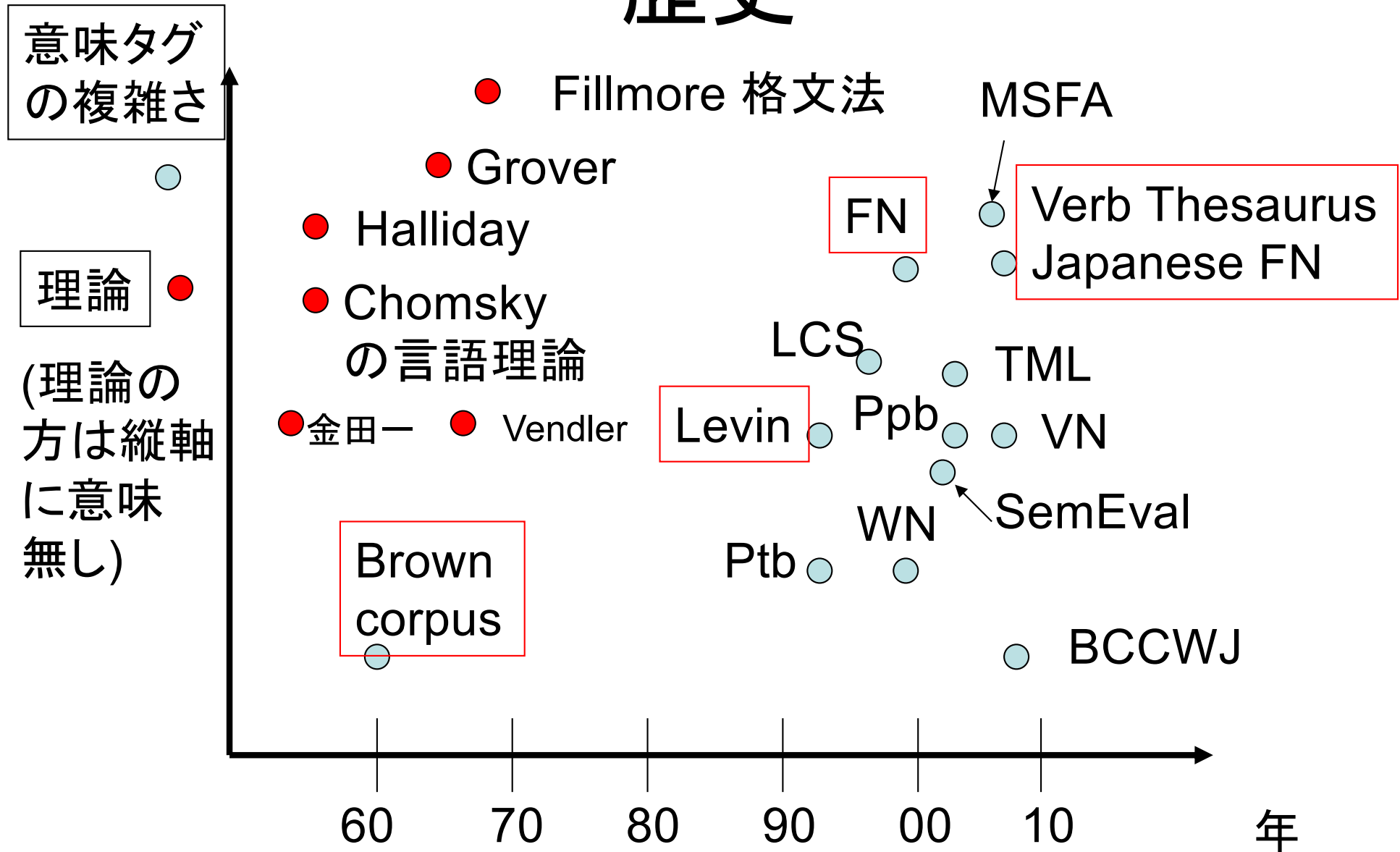
- Global Document Annotation
 - 参照, 述語項構造解析

- SUMO, (Niles and Pease, 2001)

- Suggested Upper Merged Ontology

「言語資源ポータル」に一覧あり

歴史



背景

- 言語処理

- より意味・感情・感性に近い処理を行いたい
 - 表層上にはない情報
- 分析された体系と事例が必要
- 大規模言語資源構築の失敗
 - 1つで全ての要求をまかなうのは不可能

- 言語資源の構築

- 意味タグは応用処理に依存
- 処理目標ごとに異なるタグ体系で構築
- 言語資源の散乱(Calzolari 08)

最近の中心的話題

- 言語処理の基礎

- 述語項構造に関する辞書と事例

Levinの動詞交替辞書(93)

LCS(95), VerbNet(06), FrameNet(98),

WordNet(90), TimeML(04),

動詞項構造シソーラス(08)

相互にリンク

- 言語資源の有効利用(特にヨーロッパ)

- 既存言語資源を結びつける (Language grid)

- top level の共通枠組みの提案

- 標準化 ISO への勧告

- 言語資源同士の相互リンク ((例) 約72%の精度)

- Calzolari 08

述語項構造

- 述語(動詞や形容詞, サ変名詞)の語義の曖昧性解消
 - 述語間の関係(言い換え)
 - 述語の概念スロットとの関係(フレーム|イベント)

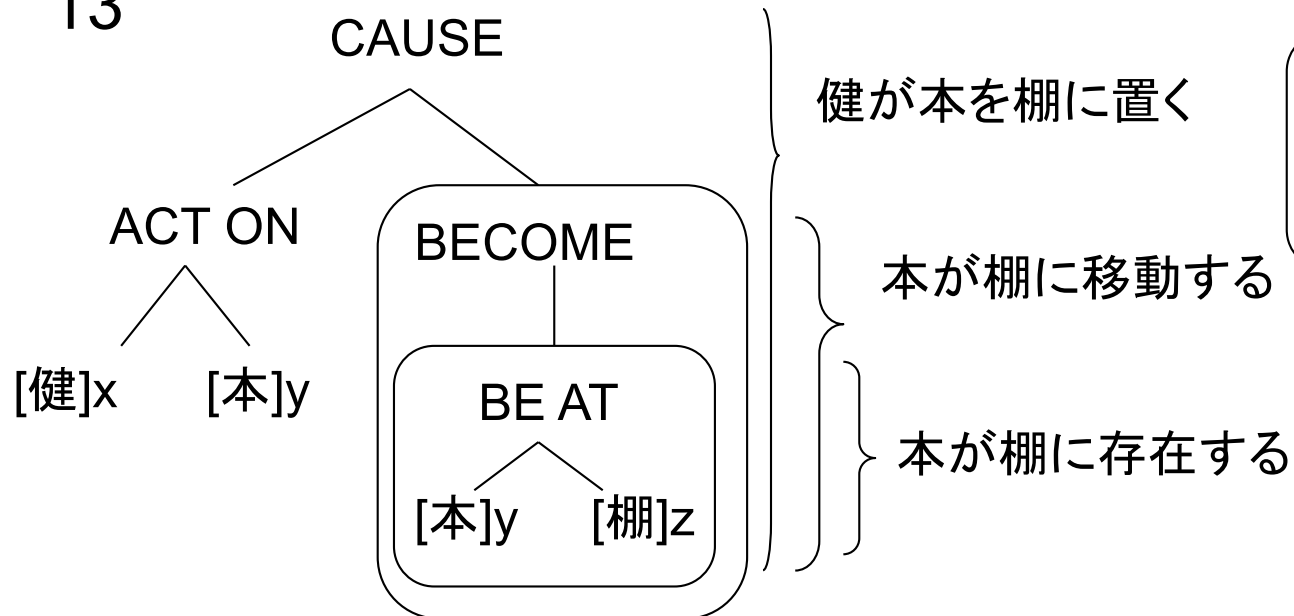
バスを 対象	1台	雇う (乗り物をお金を払って専用して使う)
-----------	----	-----------------------

会社が 動作主	税理士を 役割	雇う (お金を払って人を使う)
------------	------------	-----------------

会社が 動作主	太郎を 対象	雇う
------------	-----------	----

太郎が 動作主	会社に 着点	就職する
------------	-----------	------

13



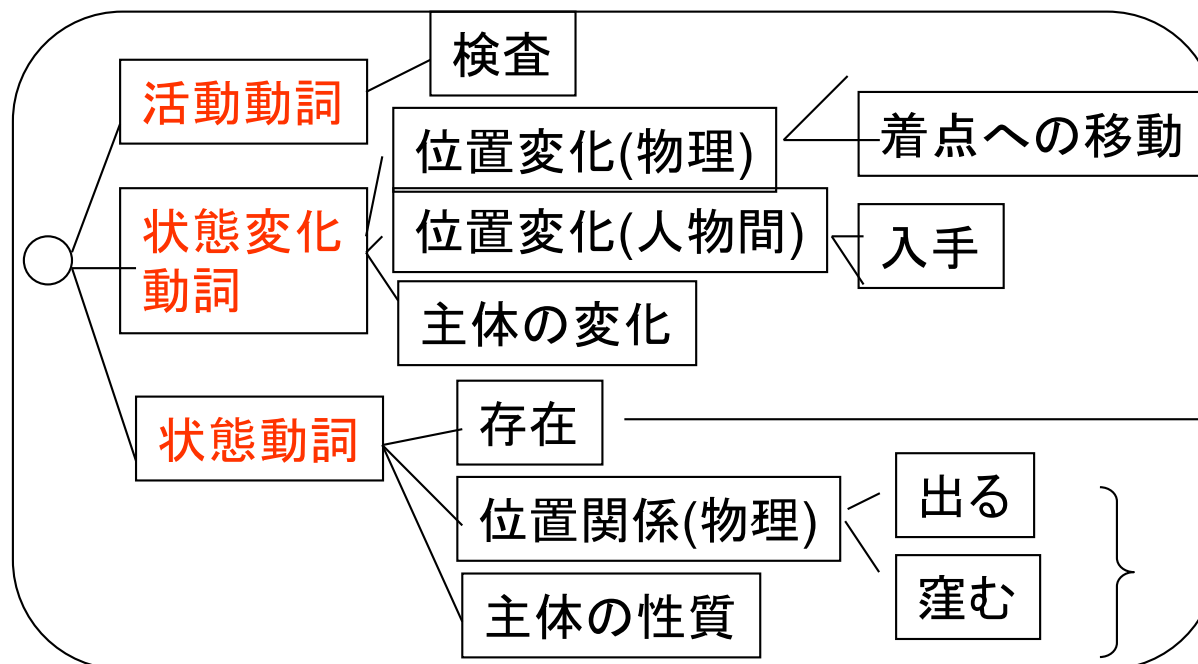
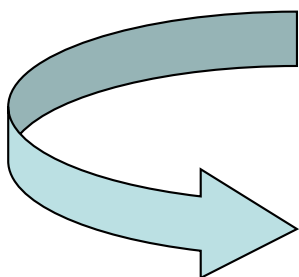
動詞項構造 シソーラス

特徴: WordNetの
ように同義語集合が
容易にとれる

上位事象

下位事象

階層的細分類

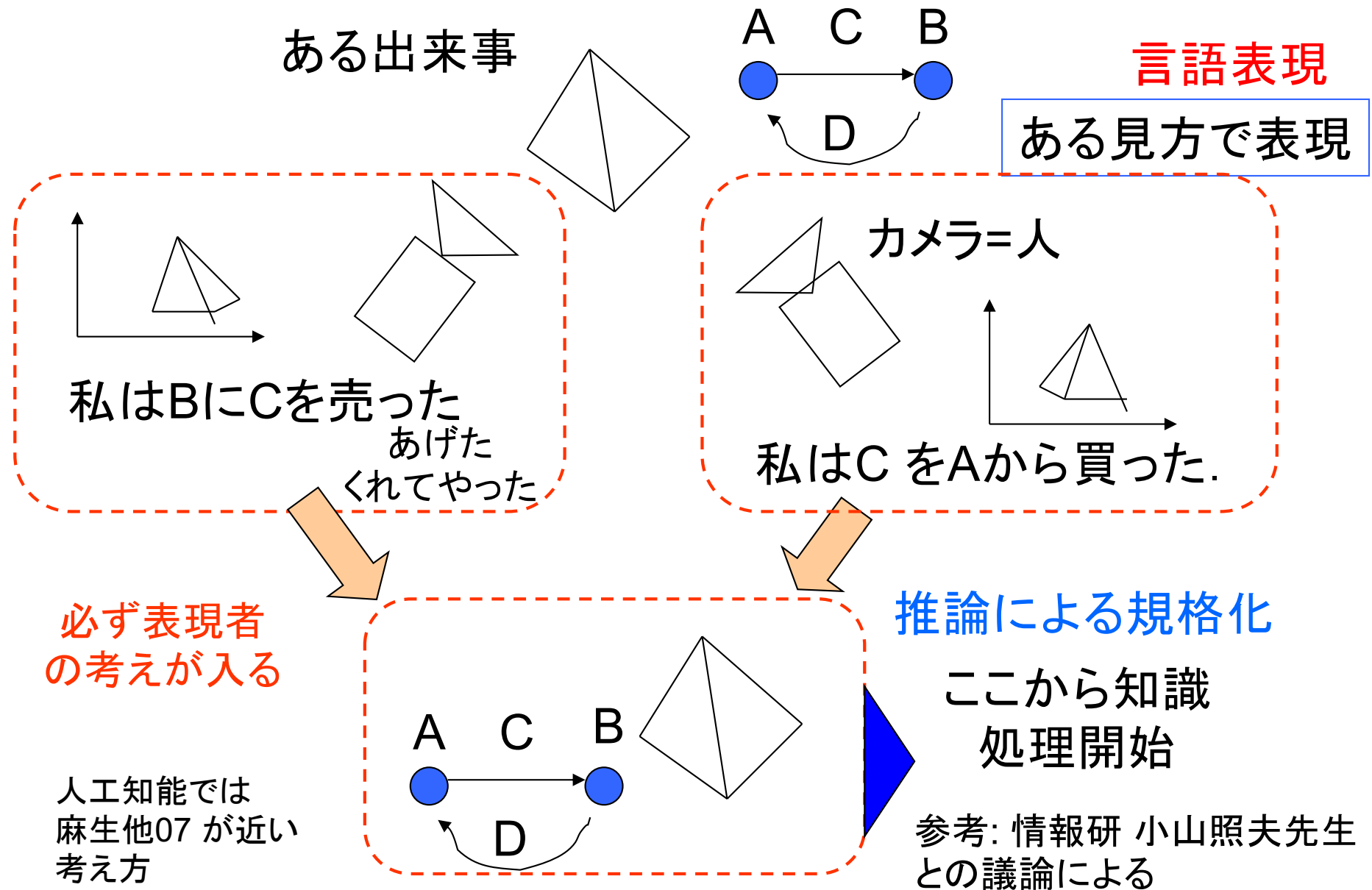


置く
移動

存在

対義関係

言語表現から内容へ：座標化



今後10年の方向

• 基礎技術

- 述語表現間の言い換えが容易になる
 - 「無線LANが壊れた/動かない/不具合」
- 単位は文を超えたevent が扱える
- 論理的な学習・推論: (決定的先見知識の利用)

• 応用に即したより深い処理

- 評判・評価情報 blog
- 文学作品の扱い
- 感情・感性
- 法令工学: 年金の回答

理論や特徴をもとにした
個別の作り込み

(北陸先端大 島津先生他)

今後10年の方向？

- オンライン言語ツールの充実
 - 翻訳者支援システム
 - 既訳文書を整理, イディオムの提示(影浦・阿辺川)
 - Yahoo! API を利用した英語校正支援
 - Native checker
 - Multiword expression の使用例
 - (例) made of のofは？

まとめ

- 言語処理のこれからの方向
 - 事態・イベントの同定と言い換え
 - Webを利用したツール
- 言語資源のこれから
 - 述語項構造資源の整備
 - 応用に特化した辞書と事例
 - 各資源間の相互リンク
 - 言語資源記述の規格化

会議・論文誌

- 会議
 - Large-Scale Knowledge Resources 2008
 - LREC conference (Lisbon-2004, Genova-2006, 2008)
 - AFNLP Asian Federation of Natural Language
 - COCOSDA (International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques)
 - SemEval
 - NTCIR (NII) などなど
- 論文誌
 - Language Resources and Evaluation など他

組織

- 組織
 - LDC (US)
 - ELRA (European Language Resource Association)
 - 言語資源協会(GSK)
 - Web日本語Nグラム
 - 岩波国語辞典第5版
 - 毎日新聞GDAコーパス

参考文献

- 麻生 英樹 高木 朗 小林 一郎 橋田 浩一. 高階表現の意味表現と述語間依存関係の解釈について. 人工知能学会年次大会, 2G5-7, 2007.
- Calzolari, Nicoletta. Initiatives, Tendencies and Driving Forces for a “Lexical Web” as Part of a “Language Infrastructure”, Proceedings of Large-Scale Knowledge Resources, LNAI 4938, Springer, pages. 90-105.
- FrameNet. <http://framenet.icsi.berkeley.edu/>
- GDA. <http://i-content.org/gda/tagguide-j.htm>
- Kuroda, Kou. MSFA他 <http://clsl.hi.h.kyoto-u.ac.jp/~kkuroda/papers.html>
- LCS: http://www.umiacs.umd.edu/~bonnie/LCS_Database_Documentation.html
- Levin, Beth. English Verb Classes and Alternations, University of Chicago Press, 1993.

参考文献

- Lexical Markup Framework rev.-14 DIS-24613 (2007), [http:// liria.fr/documents.html](http://liria.fr/documents.html)
- Native checker. <http://native-checker.com/>
- TimeML. <http://www.timeml.org/site/index.html>
- Tokunaga, T., et al.: Infrastructure for standardization of Asian language resources. In: Proceedings of COLING/ACL 2006 Main Conference Poster Sessions, Sydney, pp. 827–834 (2006)
- Takeuchi, K., Kanehira, T., Hilao, K., Abekawa, T. and Kageura, K. "Flexible automatic look-up of English idiom entries in dictionaries," *Machine Translation Summit XI Proceedings*. p. 451-458. 2007.
- VerbNet. <http://verbs.colorado.edu/~mpalmer/projects/verbnet.html>
- WordNet. <http://wordnet.princeton.edu/>