A Japanese Word Dependency Corpus
日本語の単語係り受けコーパス

森 信介　　笹田 鉄郎　　小椋 秀樹

京都大学　　京都大学　　立命館大学

2015 年 3 月 18 日

Special thanks to NTT CS 研, 国立情報学研究所, 国立国語研究所

背景

- 文の構造記述としての係り受け
- 多くの言語で単語単位が主流
  - *Bunsetsu*? What is it? と言われて辛い
  - 名詞句の構造など (文節では記述されない)
- 日本語でも単語係り受けコーパスがあるといい

  Cf. CoNLL Multilingual Dependency Parsing [Buchholz+ 2006] の日本語データは文節内は右隣
  - 大規模 (数万文, Penn Treebank [Marcus 93])
  - 多数の分野

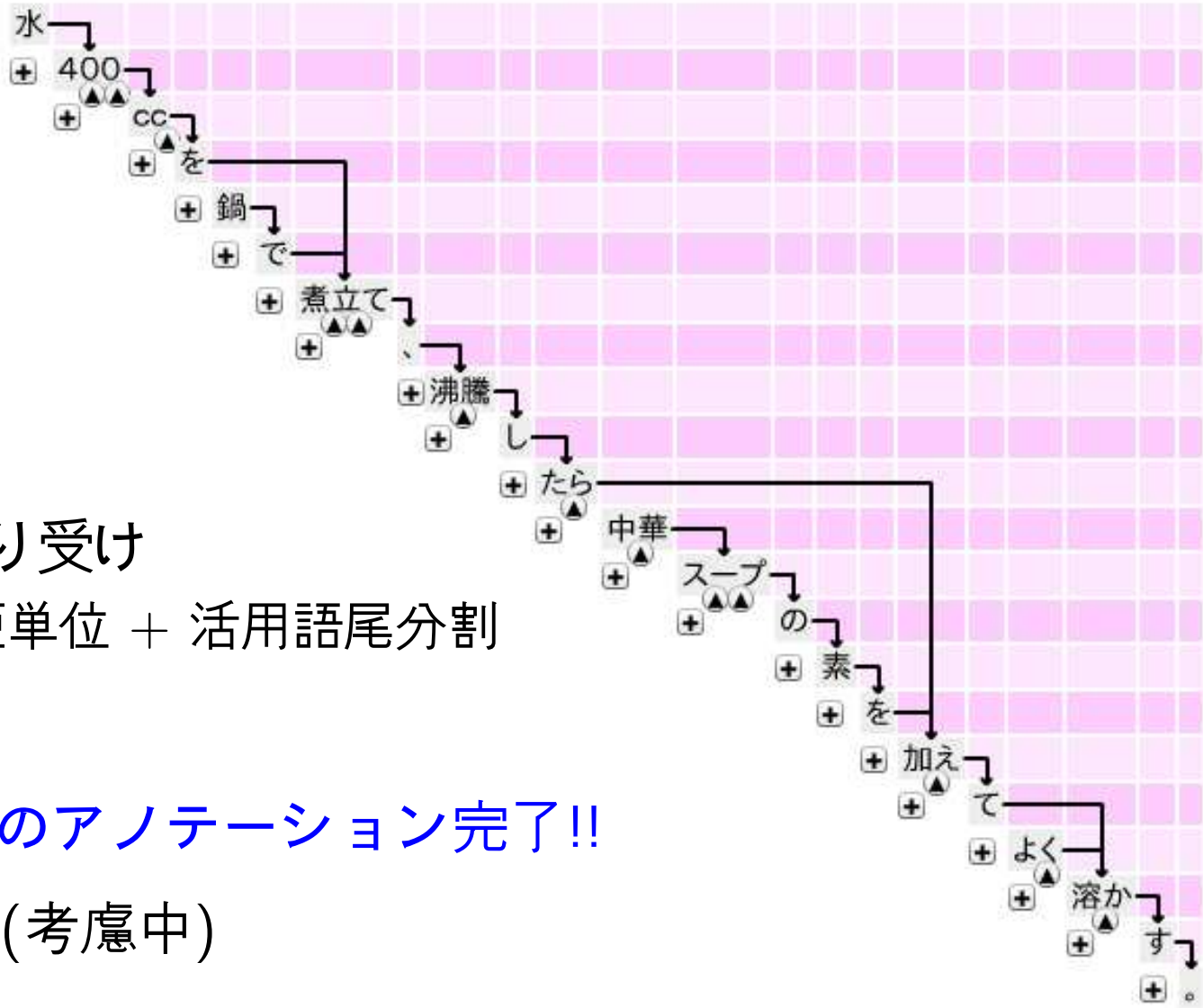
多段言語処理における解析課題の位置

1. 単語分割
2. (品詞推定)
3. (固有表現認識)
4. 係り受け解析
5. 述語項構造解析
6. ...

様々な分野のテキスト

- 『現代日本語書き言葉均衡コーパス』 (BCCWJ) [前川 09]
  - 新聞, 雑誌, 書籍, 白書, ブログ, Web QA (単語分割・品詞)
    - 世界的にみて先行
  - 様々な言語現象の共通のアノテーション対象 (Class A)
- 特許 [NTCIR 特許翻訳]
- レシピ [CookPad] (2015 年 2 月データ公開)
- 論文抄録
- 医療
- 将棋解説 (Cf. 自動解説生成 [亀甲 14])
- etc.

概要



- 単語間の係り受け
  - 単語: 短単位 + 活用語尾分割
- 交差を許容
- 約 35,000 文のアノテーション完了!!
- ラベルなし (考慮中)

アノテーション基準

- 基本的に左から右
- 主辞に掛ける
- 複数の妥当な候補があれば左側に係るとする
  例) パスタ → を → 作 → っ → て → $_{\phi\text{ヲ}}$ 食べ → る
- アノテーションの過程で作業者が判断に迷う点などを中心に基準を整備
- 著者間で最後まで意見が合わないことはなかった
  - 森 信介, 笹田 鉄郎, 小椋 秀樹 (BCCWJ 単語単位策定)
  - 初見でほぼ一致, 議論は簡単に収束

アノテーション基準

(1) 複合語: 語構造を判断 (cf. BCCWJ の短単位規程)

例) (交通→ (バリア → フリー)) → 法

(2) 複合語に係る連体修飾の係り先

- 全体に係る場合

  例) 総合 → 的 → な → (バリア → フリー → 化)

- 一部に係る場合 (単語単位の長所)

  例) 項 → 構造 → の → (曖昧 → 性) → 解消

アノテーション基準

(3) 括弧の対応

a 開き括弧は閉じ括弧に係る

例) ［ → (1 → ］)

例) 「 → (京都 → 駅→ まで→ 」)

b 括弧が付された注釈的要素の扱い

例) 国際 → (原子 → 力 → 機関) → (（ → (ＩＡＥＡ → ）)
→ の → 調査

(4) 並列

a 最後の要素に並列マーカーがない場合

例) これ → と → あれ → を

例) 衆議 → 院 → と → (参議 → 院) → が

b 最後の要素に並列マーカーがあればそれに対応付ける

例) これ → と → (あれ → と) → を

- その他多数の細かい問題

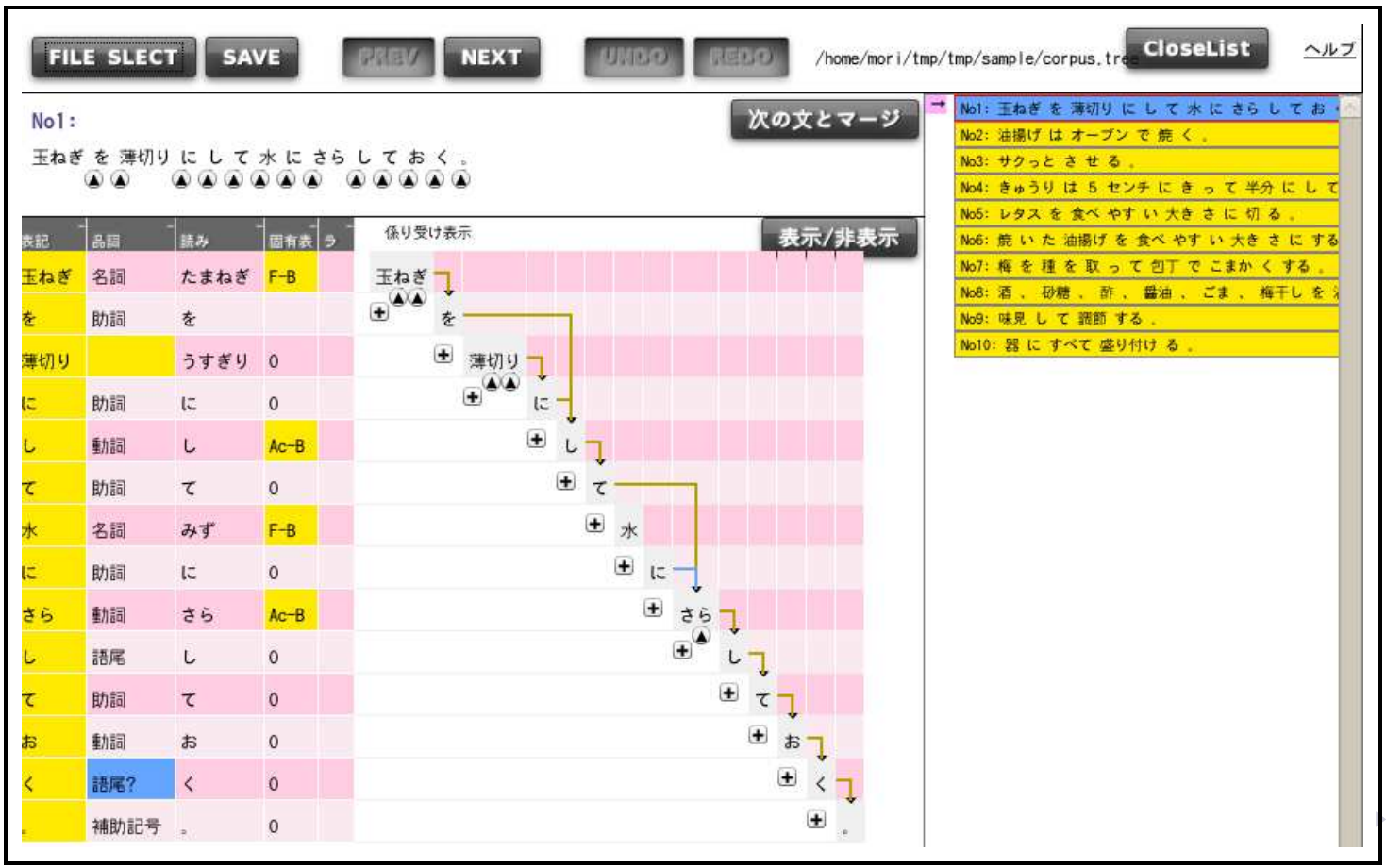
アノテーション基準

詳細は基準書を参照

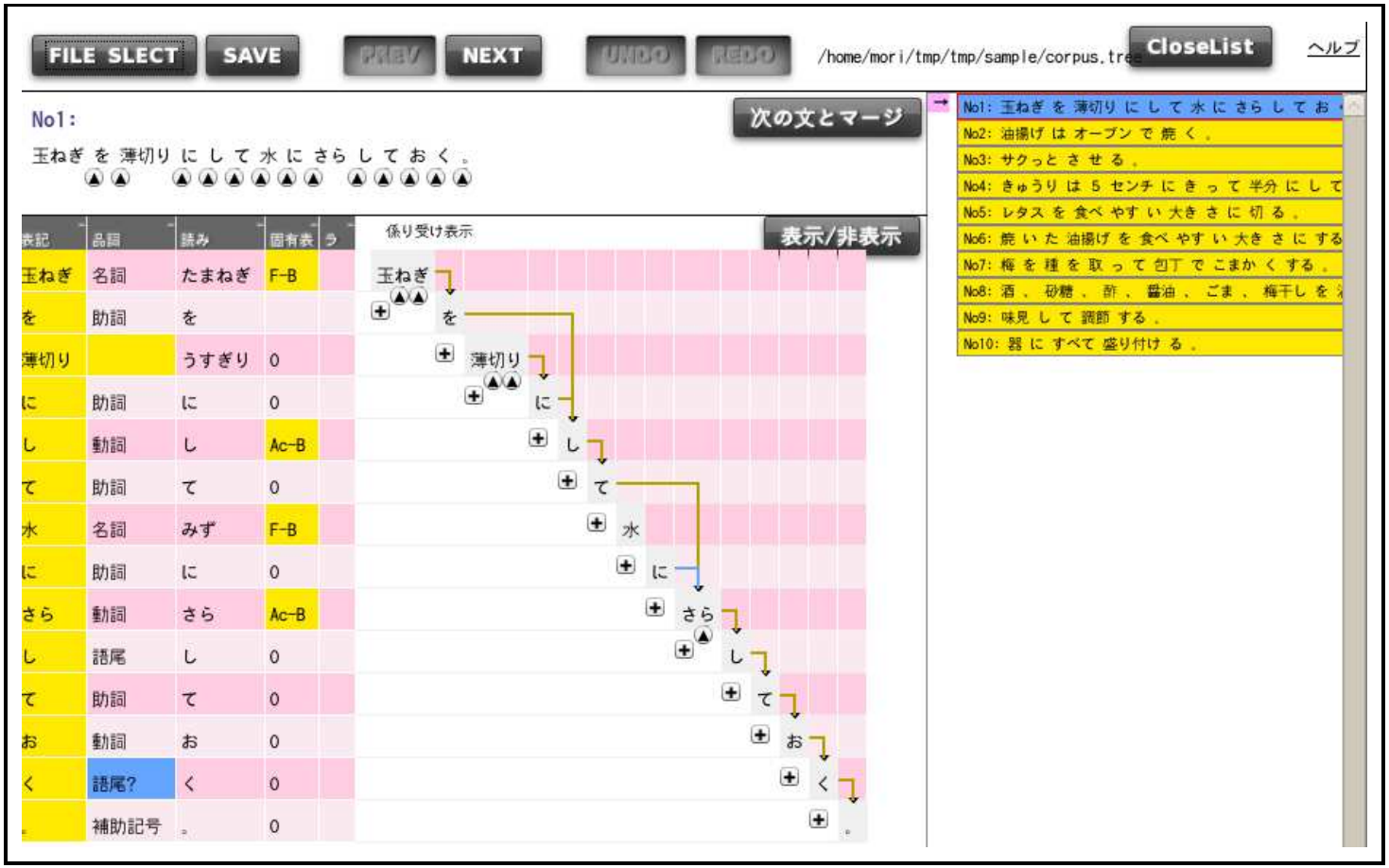
それと実例も…

PNAT: Pointwise NLP Annotation Tool

- 単語分割
- 品詞
- 読み
- 固有表現タグ
- 係り受け



PNATの特徴的機能



- チェックすべき箇所 (能動学習などのため)
- 辞書にあるか否か (初出は怪しい)
- 何らかの規則違反

諸元 フルアノテーション

- 全ての単語に係り先を付与

ID		出典	文数	単語数	文字数
BCCWJ	OC	Yahoo!知恵袋	1,115	22,333	31,047
	OW	白書	1,162	50,498	73,051
	OY	Yahoo!ブログ	1,366	22,625	33,174
	PB	書籍	1,569	35,265	48,859
	PM	雑誌	2,000	32,689	50,173
	PN	新聞	2,218	50,684	73,901
EHJ		英語表現辞典	13,000	162,397	220,146
NKN		日経新聞	10,025	292,462	442,264
NPT		NTCIR 発明開示書	2,000	81,705	127,840
JNL		論文抄録	354	13,379	22,202
RCP		レシピ	724	13,147	19,975
合計			35,533	777,184	1,142,632

諸元 部分的アノテーション

- 一部の単語にのみ係り先を付与
  - 効率的精度向上 (分野特有の語に集中，能動学習)
  - 主に学習用

ID	係り受け数	備考
KUC	294,314	京大コーパスからの自動変換
EDR	550,823	EDR コーパスからの自動変換
BCCWJ	23,000	BCCWJ (フルでカバーされない高頻度語)

部分的アノテーション戦略

1. 単語係り受けが付与されていないコーパスに出現する単語 $w \in W_{BE2}$ に対して
$$L(w) \quad = \quad \log_2 F_0(w) - F_{full}(w) - f_a(w)$$
が最大かつ $F_{full}(w) + f_a(w) < 3$ である単語 w^* を選択
   - $F_0(w)$: テストデータを除いた BCCWJ 全体での出現頻度
   - $F_{full}(w)$: 既存の学習データにおける単語アノテーション数
   - $f_a(w)$: 部分的アノテーション開始以降のアノテーション数
2. w^* の 1 つの出現箇所に対して係り先と係り先の係り先をアノテーション
3. $freq_a(w^*) + +$
4. goto 1.

部分的アノテーション戦略

- 頻度の対数値に比例して万遍なくアノテーションしたい
  - → 能動学習?
- $F(w)$ のスコア計算結果の例

回数	1st	2nd	3rd	4th	···
防災	**8.75822**	7.75822	**7.75822**	6.75822	···
裁判	7.88264	**7.88264**	6.88264	6.88264	···
未婚	6.94251	6.94251	6.94251	**6.94251**	···
⋮	⋮	⋮	⋮	⋮	⋱
w^*	防災	裁判	防災	未婚	···

- アノテーションの例
  - … 防災 → に → (長らく 携わ) り …

係り受け解析実験

- EDA [Flannery 12]
  - 部分的アノテーションから学習可能
    - 文中の一部の単語にのみ係り先を付与
  - 係り受けの交差を許容 (MST(最大全域木)ベース)
  - 条件付き探索 (固有表現の保持)
  - オープンソース
    `http://plata.ar.media.kyoto-u.ac.jp/tool/EDA/`
  - オープンリソース
    - 学習コーパス (本発表)
    - 様々な分野に対応したモデル

EDA: Easily adaptable Dependency Analyzer

- 最大全域木 [McDonald 05, McDonald 11]
- 点予測 [Neubig 10, Neubig 11]
  1. 全ての単語間の係り受けスコアを計算

$$\sigma(\langle i, d_i \rangle, \vec{w}), \qquad \text{ここで } w_i \text{ は } w_{d_i} \text{ に係る}$$

  2. エッジスコアの合計が最大になる全域木 (MST) を選択

$$\hat{\vec{d}} = \underset{\vec{d} \in D}{\text{argmax}} \sum_{i=1}^{n} \sigma(\langle i, d_i \rangle, \vec{w})$$

部分的アノテーションからも学習可能!
⇒ 自由度の高いコーパスアノテーション
⇒ 分野適応が迅速かつ安価

点予測による係り受け解析 (つづき)

- スコア計算の素性

牡蠣 を 広島 に 食べ に 行 く

w_{i-3} w_{i-2} w_{i-1} w_i w_{i+1} w_{i+2} w_{i+3}

w_{d_i-3} w_{d_i-2} w_{d_i-1} w_{d_i} w_{d_i+1} w_{d_i+2} w_{d_i+}

F1 係り元 w_i と係り先 w_{d_i} の距離
F2 w_i と w_{d_i} の表記
F3 w_i と w_{d_i} の品詞
F4 w_i と w_{d_i} の前後 3 単語の表記
F5 w_i と w_{d_i} の前後 3 単語の品詞

係り受け解析実験

- テスト コーパス
  - BCCWJ-Core PN, PM, PB, OC, OY, OW (各約 500 文)
    - Next NLP の学習・テスト の分割
  - 英語表現辞典 (1/10), 日経新聞記事 (1/10)
  - レシピ, 特許, 論文抄録
- 学習コーパス: テスト 以外のフルアノテーションすべて
  1. 部分的アノテーションなし
  2. 部分的アノテーションあり

各分野の解析精度

分野		文数	単語数	精度 [%] BCCWJ-part	
				なし	あり
BCCWJ OC	Y!知恵	500	9,846	93.10	93.18
BCCWJ OW	白書	504	23,952	88.95	89.17
BCCWJ OY	ブログ	509	9,239	92.26	92.39
BCCWJ PB	書籍	511	11,792	91.08	91.14
BCCWJ PM	雑誌	495	7,415	92.28	92.40
BCCWJ PN	新聞	505	12,621	91.24	91.28
EHJ	辞典例文	1,300	16,433	97.05	97.08
NKN	経済新聞	1,002	29,037	92.74	92.67
NPT	特許	250	10,497	93.11	93.09
JNP	論文抄録	32	1,116	92.07	92.25
RCP	レシピ	62	1,139	93.87	93.69

各分野の解析精度

分野		文数	単語数	精度 [%] BCCWJ-part	
				なし	あり
BCCWJ OC	Y!知恵	500	9,846	93.10	93.18
BCCWJ OW	白書	504	23,952	88.95	89.17
BCCWJ OY	ブログ	509	9,239	92.26	92.39
BCCWJ PB	書籍	511	11,792	91.08	91.14
BCCWJ PM	雑誌	495	7,415	92.28	92.40
BCCWJ PN	新聞	505	12,621	91.24	91.28

- 白書 (OW) が難しい (長文?)
- 部分的アノテーションの追加により精度向上

各分野の解析精度

分野		文数	単語数	精度 [%] BCCWJ-part	
				なし	あり
EHJ	辞典例文	1,300	16,433	97.05	97.08
NKN	経済新聞	1,002	29,037	92.74	92.67
NPT	特許	250	10,497	93.11	93.09
JNP	論文抄録	32	1,116	92.07	92.25
RCP	レシピ	62	1,139	93.87	93.69

- 辞書の例文 (EHJ) の精度が高い
- BCCWJ の部分的アノテーションの追加の効果はまちまち

おわりに

- 日本語の単語係り受けコーパス構築 (タグは配布可)
  - 35,000 文のフル アノテーション
  - 23,000 単語の部分的アノテーション
- 多分野&多現象 (世界的にみて先行)
  - 『現代日本語書き言葉均衡コーパス』 6 分野
    - 様々な言語現象の共通のアノテーション対象 (Class A)
  - 特許 for 機械翻訳 [Sudoh 14]
  - レシピ for 手順書の理解 [Mori 12]
  - 論文抄録, 医療
  - 将棋解説 for 自動解説生成 [亀甲 14]

今後の予定

- データ追加
- 解析精度の向上
- 文生成・自動要約への応用
- 係り受けラベル
  - 基準策定
  - アノテーション

References

- Flannery, D., Miyao, Y., Neubig, G., and Mori, S.: A Pointwise Approach to Training Dependency Parsers from Partially Annotated Corpora, *Journal of Natural Language Processing*, Vol. 19, No. 3 (2012)
- Marcus, M. P. and Santorini, B.: Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, Vol. 19, No. 2, pp. 313–330 (1993)
- McDonald, R., Pereira, F., Ribarov, K., and Hajič, J.: Non-projective Dependency Parsing Using Spanning Tree Algorithms, in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 523–530 (2005)


McDonald, R. and Nivre, J.: Analyzing and Integrating Dependency Parsers, *Computational Linguistics*, Vol. 37, No. 4, pp. 197–230 (2011)


Mori, S., Sasada, T., Yamakata, Y., and Yoshino, K.: A Machine Learning Approach to Recipe Text Processing, in *Proceedings of the 1st Cooking with Computer Workshop*, pp. 29–34 (2012)


Neubig, G. and Mori, S.: Word-based Partial Annotation for Efficient Corpus Construction, in *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (2010)


Neubig, G., Nakata, Y., and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 529–533 (2011)

- Sudoh, K., Nagata, M., Mori, S., and Kawahara, T.: Japanese-to-English Patent Translation System based on Domain-adapted Word Segmentation and Post-ordering, in *Conference of the Association for Machine Translation in the Americas, AMTA-14*, pp. 234–248 (2014)
- 亀甲 博貴, 三輪 誠, 鶴岡 慶雅, 森 信介, 近山 隆■対数線形言語モデルを用いた将棋解説文の自動生成, 情報処理学会論文誌, Vol. 55, No. 11, pp. 2431–2440 (2014)
- 前川 喜久雄■代表性を有する大規模日本語書き言葉コーパスの構築, 人工知能学会誌, Vol. 24, No. 5, pp. 616–622 (2009)