並列疑似エラー補正法に基づく「破格」な言語表現の (疑似) 解釈*
「不自然処理」のための理論的枠組み

黒田 航
京都工芸繊維大学 (非常勤)　早稲田大学総合研究機構情報教育研究所 (客員研究員)
kow.kuroda@gmail.com

1　はじめに

1.1　「不自然言語処理」の部分問題

異表記，誤表記，新規表現，誤用や逸脱から生じる用法を総称して破格表現と呼ぶ[1]．新表記と新語，逸脱用法の実例を，下に幾つか挙げる:

(1) 新表記や新語の例[2]
  a. ネ申 (「神」の新表記), 儲 (「信者」の新表記),
  b. ようつべ (“YouTube” の新表記), ガイシュツ (「既出」の新表記), マンセー (「万歳」の新表記), ふいんき (← 何故か変換できない）, ちょwwwおまwwww」(「ちょっとお前，何言ってんの？」の異形)
  c. 氏ね (「死ね」の新表記), 人大杉 (「人多すぎ」の新表記), 空気嫁 (「空気読め」の異表記?), マスゴミ (「マスコミ」の異表記?), 腐女子 (「婦女子」の異表記?), 全俺が泣いた (誇張表現「全米が泣いた」のパロディー)

新表記と新語の境界は非常に曖昧である．

(2) うろ覚え (誤用) や逸脱で起こる用法の拡張[3]
  a. 時間をもてあそぶ
  b. 熱血感, 門外感, 衰弱さ, 唖然さ

これらの Web 文書に多発する破格表現は NLP の鬼門である．その理由は次の通り:

(3) a. 破格表現はしばしば辞書に存在しない要素を含み，有効な形態素解析結果が得られない.
  b. 破格表現はしばしば文法的に正しくない表現であり，有効な構文解析結果が得られない.

(3) に挙げた問題は以前から指摘されてきた．だが，現時点で有効な対処法があるようには見えない．それは標準的な処理モデルで，形態素解析が「閉じた辞書」を前提にし，構文解析が「閉じた文法」を前提にしているからである．

「閉じた辞書」や「閉じた文法」の想定が現実に則していないことは，次の考察から間接的に明らかになる: (1)–(2) に挙げたような表現はどうやって生成されるのか? [13] によれば，これらは本来の意味で「生成」されたものではなく，慣習化された既成表現 (prefabricated sequences) が追加編集を伴って「再生産」されたものである[4]．これと整合するのは，記憶ベースで類推ベースの言語処理であるが，その実装は，NLP 研究者が自らの「言語の見方」を変えないと実現できないように思える．

1.2　本論文の提案

この問題意識から本発表が目指すのは，膨大な事例記憶 [11, 14] の想定に基づいて「閉じた辞書」と「閉じた文法」を前提にしない事例ベースの言語処理システムの概略を示すことである．具体的には，並列疑似エラー補正 (Parallel Simulated Error Corrections: PSEC) の結果の統合と名づけたモデルの「不自然言語処理」への応用例を紹介する．

2　並列疑似エラー補正

並列疑似エラー補正 (PSEC) は (4) で [12, 14] で定義された，次のような事例ベースの処理である:

(4) 入力が n 個の部分 $x_1, x_2, \ldots, x_n$ からなる系列だとする．中立性を確保するために x_i を X の i 番目の文節と呼ぶ.
  a. 初期化:　i 番目 ($i \in n$) の文節 (e.g., x_i) を変項化した状態 $X'_1, X'_2, \ldots, X'_n$ を仮想的に作り出す．この状態は表 1 で表現される状態である (Δ が変項化した箇所を表わす)．この時，X は $\{X'_1, X'_2, \ldots, X'_n\}$ の重ね合わせ (superposition) = 構造の単一化と定義される．
  b. 疑似エラー補正による x_i の補完:　X'_i ごとに変項 Δ_i の値の補完を行なう (これが疑似エラー補正 (simulated error correction) と呼ぶ処理)．この補正処理の結果を表 2 で表わす (x''_i が Δ_i で補完された新しい値で，X''_i は X'_i の Δ_i の値が補完された疑似的な実例である).
  c. (4b) による補完のための (有効な) 候補の数が 0 だった場合には，導入する変項の数を一つ増やし，探索範囲を広げる．一般に探索範囲はパターン束 (pattern lattice) [12, 14] で与えられる (e.g., 図 3).
  d. 補完結果の統合:　$X''_1, X''_2, \ldots, X''_n$ を重ね合わせ/単一化によって統合する．ただし，統合の際に不一致が起る場合と起こらない場合で，処理の難易度が異なる．不一致が生じる場合は，選択が必要になる．

$X = x_1 \cdot x_2 \cdot x_3 \cdot x_4 \cdot x_5$ のパターン束を図 1 に示す．この図の右端が実例 X であり，右から二列目の 5 つのパターンが表 1 に対応する (図中の ＿ が表中の Δ に対応).

(4b) の疑似エラー補正処理は技術的な重要な課題となる．本稿では，膨大に事例記憶 [11, 14] に基づく事例ベースの類推 [2, 8, 17] によって実現すると考える．このような処理システムは一般に実装が困難であるが，PSEC の効能は，疑似エラー補正が並列分散化され，その稼働性が向上している点にある．

以下では分析例を示しつつ PSEC の有効性を述べる．

3　疑似エラー補正の実例と応用例

3.1　超語彙的パターンによる構文効果とブレンド効果の記述

PSEC は次の表現の解釈特性を記述するために考案された:

(5) a. その絵は壁にかかっていた.
  b. その男は医者にかかっていた.
(6) a. ?*その絵は医者にかかっていた.
  b. ?*その男は壁にかかっていた.

* 小松原哲太 (京都大学大学院) と長谷部陽一郎 (同志社大学) から本稿執筆過程で受けたコメントに感謝する.

1) Web 文書にこれらに多発する理由の一部は，[13] が論じているように，従来の出版物を構成する書き言葉が出版の過程で標準化のための編集 (一種の検閲) を受けるのに対し，Web 文書には標準化のための編集がほとんど働かないことに求められる.

2) このクラスの実例は [10] に基づいている.

3) このクラスの実例は [13] に基づいている.

4) この時，利用可能な表現の集合は一種の「公共財」となっている．この財の使用には母語話者性 [7] のシグナル効果 [9] がある.

表 1　$X = x_1 \cdot x_2 \cdots x_n$ の PSEC の初期状態 (=1 個の疑似エラーが発生した n 通りの状態)

X	x_1	x_2	$\cdots$	x_i	$\cdots$	x_n	変項化
X'_1	Δ_1	x_2	$\cdots$	x_i	$\cdots$	x_n	$x_1 \Rightarrow \Delta_1$
X'_2	x_1	Δ_2	$\cdots$	x_i	$\cdots$	x_n	$x_2 \Rightarrow \Delta_2$
X'_i	x_1	x_2	$\cdots$	Δ_i	$\cdots$	x_n	$x_i \Rightarrow \Delta_i$
X'_n	x_1	x_2	$\cdots$	x_i	$\cdots$	Δ_n	$x_n \Rightarrow \Delta_n$

表 2　$X = x_1 \cdot x_2 \cdots x_n$ の疑似エラー Δ_i の補正後の状態 (x''_i は特定の語句)

X	x_1	x_2	$\cdots$	x_i	$\cdots$	x_n	補正
X''_1	x''_1	x_2	$\cdots$	x_i	$\cdots$	x_n	$\Delta_1 \Rightarrow x''_1$
X''_2	x_1	x''_2	$\cdots$	x_i	$\cdots$	x_n	$\Delta_2 \Rightarrow x''_2$
X''_i	x_1	x_2	$\cdots$	x''_i	$\cdots$	x_n	$\Delta_i \Rightarrow x''_i$
X''_n	x_1	x_2	$\cdots$	x_i	$\cdots$	x''_n	$\Delta_n \Rightarrow x''_n$

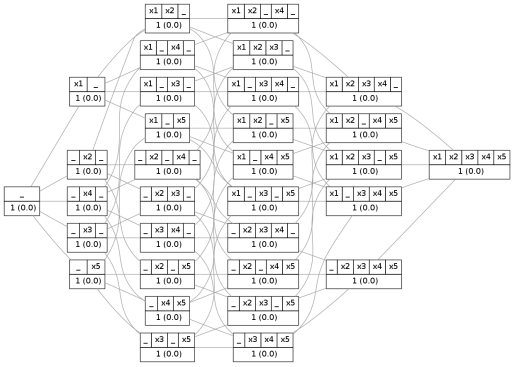

図 1　$X = x_1 \cdot x_2 \cdot x_3 \cdot x_4 \cdot x_5$ のパターン束 ("_" が変項を表わす)

(5) と (6) の比較から，⟨吊り下がり⟩ (part-of ⟨展示⟩) の状況や ⟨(不) 定期受診療⟩ の喚起は，「その絵 (は)」「その男 (は)」「壁に」「医者に」「かかっていた」という個々の句の意味の合成による効果ではなく，(7) と (8) の超語彙的パターンがフレーム喚起の効果をもつからだと考えるとうまく記述できる:

(7)　a. その N_1 は壁にかかっていた
　　b. その絵は N_2 にかかっていた
(8)　a. その N_3 は医者にかかっていた
　　b. その男は N_4 にかかっていた

ただし $N_1, N_2, \ldots, N_3$ は異なる実現の期待値をもち，解釈バイアスが生じる．これを (7) と (8) の超語彙的パターンによる特定のフレーム喚起の効果であると定義すれば，構文効果 [4] やブレンド効果 [3] が自然に説明される．

3.2　新語の意味解釈

3.2.1　無意味語句の意味評定

疑似エラー補正にどれぐらい実在性が期待できるかを調べた研究に [6] がある．この研究は (9) から (10) に挙げた文を被験者に提示し，それらを意味特徴を評定するように求めた．結果は，具体的な状況との対応づけから無意味語に有意味な解釈が生成されることが示している．

(9)　a. 通り魔が下校中の小学生を襲った．
　　b. 通り魔がユチを襲った．
　　c. ツサが下校中の小学生を襲った．

(10)　a. 大型の台風が日本列島を襲った．
　　b. 大型の台風がホヌを襲った．
　　c. レウが日本列島を襲った．

個々の事例の解釈で，どんな事例からの類推が働くのかは，図 2 に示したパターン束から読み取ることができる (パターン束の詳細は [5, 14] を参照).

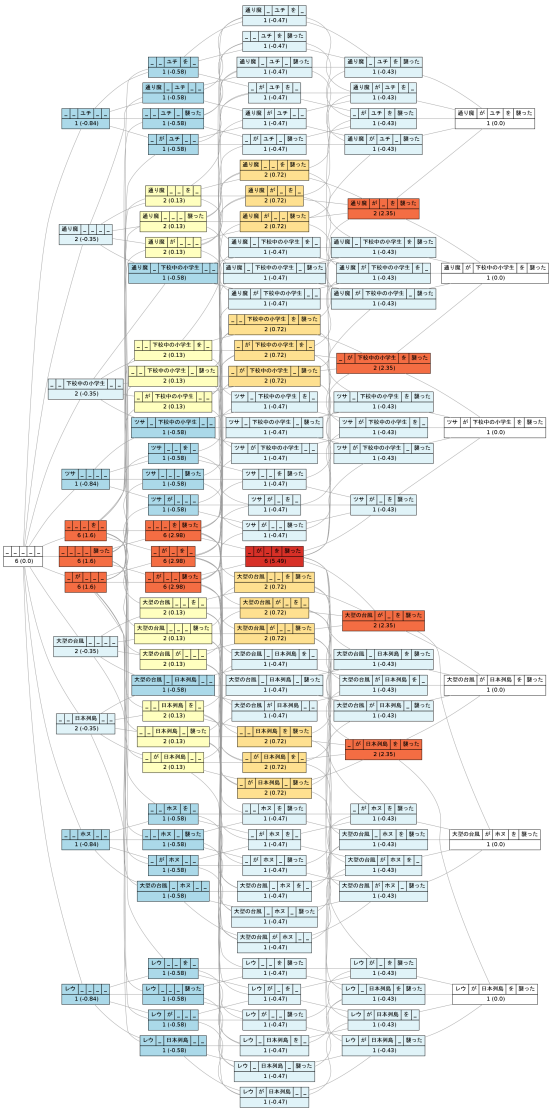

図 2　(9) と (10) のパターン束 ("_" が変項を表わす)

例えば (9b) は (9a) との顕著な類似性により，「ユチ」が「小学生」の同類と解釈されるバイアスが存在する．これは「ユチ」と「小学生」がいずれも「通り魔が Δ を襲った」の実例になっていることによる．同様に (9c) は (9a) との顕著な類似性により，「ツサ」が「通り魔」の同類と解釈されるバイアスが存在する．これは「ツサ」と「通り魔」がいずれも「Δ が下校中の小学生を襲った」の実例になっていることによる．

「ユチ」や「ツサ」などの無意味語句が有意味に解釈可能であるという結果は，解釈が必要となる語句 x_i の意味が不明でも，その生起文脈の意味が十分に特定的であれば，推測可能であることを意味している．これは新語の意味解釈に応用できる．

とはいえ，この説明が意味をもつには次のような条件がある:

(11)　a. 補完へのバイアス強度の異なり

b. 文節化の事前知識

それぞれについて，以下で詳しく論じる．

3.2.2 補完へのバイアス強度の異なり

十分に有意味な補完が可能になる疑似エラーとそうでない疑似エラーの違いが存在する．具体的に言えば，(9b) の解釈で，「通り魔が Δ を襲った」は有意味な補完を可能にする疑似エラーだが，他の疑似エラー (i.e.,「Δ がユチを襲った」「通り魔 Δ ユチを襲った」「通り魔がユチ Δ 襲った」「通り魔がユチを Δ」) はそれと同じ程度に収束性のある補完が可能でない．

一般にパターン p のバイアス強度は，他の条件が同じであれば p の生産性の度合い (図 2 では色温度で表わされている) で測ることができると考えられる (ただ生産性の指標の選び方に課題が残されている)．他に考慮すべき条件とは，語義の類似性である (「死人に口なし」と「死者に口なし」は単に編集距離の近さによって類似性が高くなっているわけではない)．今の時点で変項の特定の値の実現に語義がどう関わっているかはそれほど明らかではないが，幾つかの特徴を §3.4 で述べる．

表現 e が少なくとも一つの強い補完を起こす超語彙的パターンを実現していれば，e の解釈が収束するには十分である．逆に，表現 e が二つ以上の強い補完を起こす超語彙的パターンを実現している場合，e の解釈は収束しないか，少なくとも収束に時間がかる．最初に挙げた (6) の解釈でそれが起こる．

3.2.3 文節化の事前知識

文節化のための事前知識なしでは図 2 に示したパターン束は実現できない．具体的に言えば，図 2 のパターン束の生成のために `Pattern Lattice Builder` (PLB)[5]に与えられた入力は (12) と (13) である:

(12) a. (通り魔, が, 下校中の小学生, を, 襲った)
b. (通り魔, が, ユチ, を, 襲った)
c. (ツサ, が, 下校中の小学生, を, 襲った)

(13) a. (大型の台風, が, 日本列島, を, 襲った)
b. (大型の台風, が, ホヌ, を, 襲った)
c. (レウ, が, 日本列島, を, 襲った)

$X = (x_1, x_2, \ldots, x_n)$ は X が n 個の要素 $x_1, x_2, \ldots, x_n$ の系列であることを表わしている．

図 2 で類似例の対応づけがうまく行っているのは，パターン束の生成の際に文節数を 5 つに揃えているからである．この種の事前知識を与える方法は自明ではない．だが，§3.5 で述べる理由で，従来の構文解析はその要求に応えられない．

3.3 パロディーの検出と意味解釈

ヒトは定型的/慣用的な表現を創造的に使う．その上，ヒトが自分の発話でパロディーを実行するのは稀なことではない．パロディーには重要なコミュニケーション上の機能がある[6]．

(14) a. 学問に王道なし
b. 死人に口なし
c. (X の) 看板に偽りなし
d. 触らぬ神に祟りなし

Web 上で検索すると，これらの諺のパロディー (あるいはうろ覚えによる誤用) が数多く見つかる．例えば (14) で挙げた諺のそれぞれに，(15)–(18) に挙げるような変異形が見つかる:[7]

(15) (14a) のパロディー
a. {i. ダイエット; ii. 外国語学習; iii. 相場; vi. 婚活; v. 脆弱性対策; vi. 物件選び} に王道なし
b. 学問に {i. 近道; ii. 横道; iii. 国境} なし
c. 学問に抜け道あり

(16) (14b) のパロディー
a. {i. 主任; ii. 模型; iii. 墜落機} に口なし
b. 死者に口なし

(17) (14c) のパロディー
a. {i. 評判; ii. 視聴率; iii. キャッチコピー} に偽りなし
b. 看板に偽りあり

(18) (14d) のパロディー
a. 触らぬ {i. ブログ; ii. 姑; iii. クレーマー} に祟りなし
b. 下らぬ株に祟りなし
c. 触れる神に祟りあり

PSEC 法はパロディーの検出と意味解釈に役に立つ．表現 e' の疑似エラー補正の結果が e' の「原典」表現 e と同一であることが，e' が e のパロディーの定義そのものだからである．

(14) と (15)–(18) の代表例からなる事例集合のパターン束を図 3 に示す[8]．従来の NLP の技術では，この種のパロディーの検出は不可能に近いが，PSEC ではそれが可能である．

(15)–(18) に挙げたような表現が (14) の表現のパロディーだとわかる条件は，(a) 聞き手が原典表現を覚えていて，(b) それらが容易に参照可能な状態にあることである．後者の条件の明示化が必要である．(15)–(18) の表現の原典参照は，「__ に口なし」「__ に王道なし」「__ に偽りなし」「看板に偽り __」「触らぬ神 __ に祟りなし」のような，説明力の高い (=色温度の高い) パターンが存在し，原典表現との結びつけを実現しているからである (説明力の高さには編集距離の短さが必要だが，それで十分ではない)．

3.4 編集の類型

原典表現 O のパロディー P の編集を対 $e = (o, p)$ で表わす (o は原典中の語句，p はパロディー中の語句)．例えば (14a) を O，(15a) を P とすると，e = (学問, ダイエット) である．

観察データから (19) のような e の類型化が可能である:[9]

(19) a. o と p が同義語: e.g., (14b) で e =(死人, 死者))
b. p が o の下位語: e.g., (14a) で e =(学問, 外国語学習); (14d) で e =(神, クレーマー)[10]
c. p と o が同類語 (=兄弟語): e.g., (14b) で e =(看板, 評判), (学問, ダイエット)
d. p が o の対義語: e.g., (14b) で e =(王道, 近道); (14c) で e =(なし, あり)
e. p が o の語呂合わせ: (14b) で e =(死人, 主任)

NLP にとって厄介なのは，パロディーの実例がシソーラスのような語彙資源を使えば対処可能な (19a) や (19d) の場合ばかりではないという点である．

(19c) の場合は，o と p の類似性が高い場合と低い場合が存在する．(15a) の e =(学問, ダイエット) は類似度が低い場合である．この場合，編集を媒介するのは抽象的な意味となる．これは特に自動 NLP が難しい場合だろう．

3.5 意味解析に有効な構文解析とは?

NLP 研究者の多くは意味解釈に形態素解析と構文解析が必要だと考える．前者は妥当な想定だと言えるが，後者については疑問の余地がある．結果に句構造を返すような構文解析は情報損失が多過ぎる．依存構造解析でも [5] が指摘した理由で情報損失は無視できない．

ヒトがパロディーを驚くべき効率と精度で処理できるという事実は，ヒトが事例ベースの言語処理をしていることを強く示

---

[5] `http://www.kotonoba.net/rubyfca/` で公開中.

[6] 定型性のある慣用表現の「創造」的な (再) 利用の研究は [15, 16] を参照.

[7] データを見るとわかることだが，パロディーの可能性は意外と活用されていない．例えば Google 検索では「に王道なし」は見つからなかった.

[8] 文節数を揃えるため，「N に口なし」「N に王道なし」「N に抜け道あり」「N に偽り {あり; なし}」の先頭にダミー文節 X を追加している．これは明らかに後知恵である.

[9] (19) に挙げた 4 つの場合でうろ覚えで生成されやすいのは (19a) である.

[10] ただし，この例では「クレーマー」は「神」の下位語ではなく，「神」のアドホック概念 [1] の ⟨逆らいえない絶対者⟩ の下位語になっている.

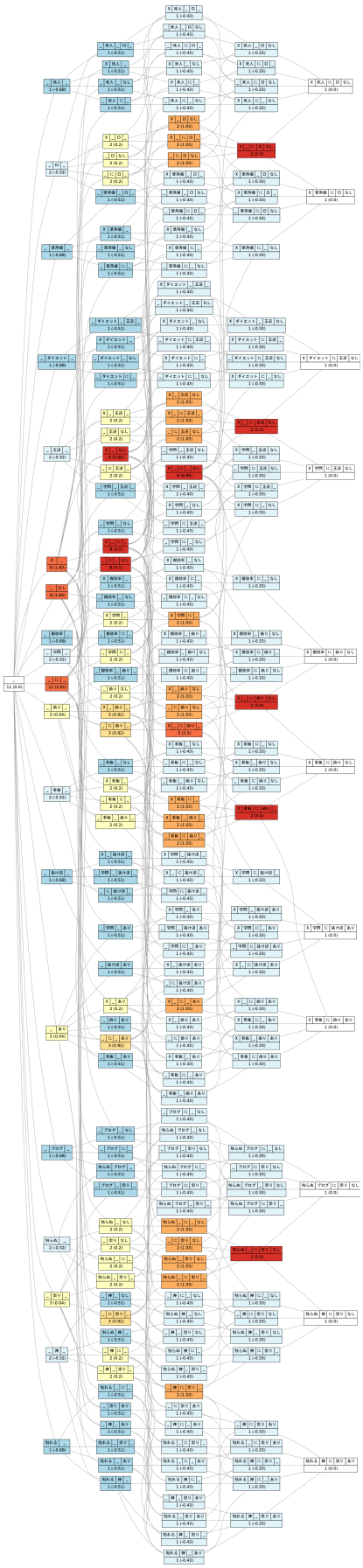

図 3　パターン束 (“_” が変項を表わす)

唆している [2, 8, 17]．パターン束の有効性はパロディーの処理に限られないので，事例ベースの言語処理が必要とする構文解析は，パターン束であると主張できる[11]．

4　終わりに

本論文は並列疑似エラー補正 (Simulated Parallel Error Correction) という手法を紹介し，それが非文法的な表現を含めた「破格な表現」の一部をうまく扱えることを示した．それにより，PSEC が新語 (や誤用) とパロディーの処理を含む「不自然言語処理」の課題の解決に貢献しうることを本発表は示した．

参考文献

[1] Laurence W. Barsalou. Ad hoc categories. *Memory & Cognition*, 11:211–227, 1982.

[2] Walter Daelemans and Antal van den Bosch. *Memory-based Natural Language Processing*. Cambridge Unversity Press, 2005.

[3] Gilles R. Fauconnier. *Mappings in Thought and Language*. Cambridge University Press, 1997.

[4] Adele D. Goldberg. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, 1995.

[5] Kow Kuroda. Arguments for *Parallel Distributed Parsing*: Toward the integration of lexical and sublexical (semantic) parsings. In *Proceedings of the 24th PACLIC*, pp. 455–462, 2010.

[6] Kow Kuroda, Keiko Nakamoto, and Hitoshi Isahara. When nonce words behave like “real” words: A case study of the Japanese verb *oso(warer)u*. In P. Bouillon and K. Kanzaki, editors, *Proc. of the 4th Intern. Conf. on Generative Approaches to the Lexicon*, 2007.

[7] Andrew Pawley and Frances Hodegtts Syder. Two puzzles for linguistic theory. In J. Richards and R. Smith, editors, *Language and Communication*, pp. 191–226. Longmans, New York, 1983.

[8] Royal Skousen. *Analogical Modeling of Language*. Kluwer Academic Publisher, 1989.

[9] Michael Spence. Job market signaling. *Quarterly Journal of Economics*, 87(3):355–374, 1973.

[10] 黒田 一平. インターネットスラングの認知言語学的考察: 卒業論文にむけて 2, 2010. 京都大学総合人間学部「言語フォーラム」(2010/12/16) で用いた発表資料.

[11] 黒田 航. 徹底した用法基盤主義の下での文法獲得: 「極端に豊かな事例記憶」の仮説で描く新しい筋書き. **月刊言語**, 36(11):26–34, 2007.

[12] 黒田 航. パターンのラティス下での疑似並列エラー修復に基づく文意の構築. In **日本認知科学会第** *26* **回大会発表論文集**, pp. 236–237, 2009.

[13] 黒田 航 and 寺崎 知之. 言語の「自然態」を捉える言語理論の必要性. In **言語処理学会第** *16* **回年次大会発表論文集**, 2010.

[14] 黒田 航 and 長谷部 陽一郎. Pattern Lattice を使った (ヒトの) 言語知識と処理のモデル化. In **言語処理学会第** *15* **回大会発表論文集**, pp. 670–673, 2009.

[15] 土屋 智行. 言語の創造性の基盤としての定型表現: 慣用句およびことわざの拡張用法の調査. In **日本認知科学会第** *27* **回大会発表論文集**, P2–19, 2010.

[16] 土屋 智行. 定型から逸脱した言語表現の分析. In **第** *17* **回言語処理学会大会発表論文集**, 2011.

[17] 佐藤 理史. **アナロジーによる機械翻訳**. 共立出版, 1997.

[11] パターン束に固有の限界がある．その一つが計算量の増加である．