

2-6 より深層的な自然言語処理のための言語資源の開発

2-6 *Development of Language Resources for Natural Language Processing in Deep Level*

張 玉潔 黒田 航 和泉絵美 野澤 元

ZHANG Yujie, KURODA Kow, IZUMI Emi, and NOZAWA Hajime

要旨

自然言語処理の基本技術として、テキストあるいは音声を書き起こしたテキストに対する分析技術は依然性能が高くない。問題は、人間の内省により規則を作ることに限界があり、また言語知識を自動抽出する場合には、浅い情報タグ付きコーパスあるいは生コーパスから深いレベルの知識を得ることが不可能なことである。解決手法として、言語材料が集まっているコーパスに対し、その用途に応じて深いレベルの情報タグを付与することが考えられる。NICTの自然言語グループでは深層的な言語処理のため言語資源を開発している。この論文では、その中の意味情報付きコーパス、学習者言語のコーパス及び日中対訳コーパスについて紹介する。

Techniques for both text analysis and speech transcription are still in an unsatisfactory state even though they cannot be dispensed with Natural Language Processing. To achieve high performance techniques, a large amount of language resources are urgently required. To resolve this problem, Computational Linguistics Group at NICT has been constructing language resources of several sorts, targeting different cases of application. This paper presents some of such resources, including Corpus Annotated for Semantic Frames and their Elements, Japanese Learner's Corpus, and Japanese-Chinese Parallel Corpus.

[キーワード]

自然言語処理, 複層意味フレーム分析, 複層相互行為分析, 学習者コーパス, 誤り分析, 対訳コーパス, 翻訳辞書

Natural language processing, Multi-planar/Multi-granular Semantic Frame Analysis, Semantic annotation of texts and dialogues, Learner corpus, Error analysis, Parallel corpora, Translation dictionary

1 まえがき

自然言語処理はコンピュータが人間の言語処理能力あるいはその一部分を実現できるように、数学モデルを考案し言語知識を構築する研究である。言語知識の獲得は、人間の内省により規則を作成する手法と、統計手法によりテキストあるいは音声データのコーパスから自動抽出する手法がある。コーパスは真実の言語材料であり参考の元になるが、その中に隠れている言語知識と言語現象を明確化するためには、コーパスに様々な側面からタグ付与する必要がある。コーパスに多い情

報を付与すればするほど、そこから多い知識を得ることができる、一般的に認識されている。以下の**2**、**3**、**4**では、NICTの自然言語グループで開発している意味情報付きコーパス、学習者言語のコーパス及び日中対訳コーパスについてそれぞれ紹介する。

2 MSFA/MIFA を使った書きコトバ/話しコトバの意味資源開発 (黒田・野澤)

言語資源開発は辞書開発に始まり、タグつきコーパスの開発と発展してきた。現状でもタ品詞情報、構文情報がタグづけられたものコーパスはそれなりの数が存在するが、技術上の困難もあって、意味情報がタグづけられたコーパスはほとんど存在しない。私たちが目指しているのはこのギャップを埋めることである。複層意味フレーム分析 (Multilayered/Multigranularity Semantic Frame Analysis : MSFA) [2][3] と、その対話構造解析への応用版である複層相互行為フレーム分析 (Multiplanar/Multigranular Interactional Frame Analysis : MIFA) [4] を使ったタグづけは意味タグづけの試みである。両者の違いは対象となる文章の違いである。MSFA はどちらかと言うと書

きコトバに特化し、MIFA は話しコトバに特化している。以下、黒田が MSFA を使った書きコトバ部門を、野澤が MIFA を使った話しコトバ部門の活動をそれぞれ解説する。

2.1 MSFA を使った書きコトバの意味資源開発 (黒田)

書きコトバ部門の活動の目的は、コーパスから採集した具体事例に、言語学者が人手で多次元の意味解析、タグづけを行い、(辞書とは別の形の) 言語資源の構築を行うと同時に、その成果を「見本」として順次、一般公開することである。この意味解析には次の特徴がある：(i) Berkeley FrameNet [1] のタグづけ仕様を改良した MSFA という解析手法を使って、意味フレームとその要素 (= 意味役割) を意味タグとして付与する。(ii) 意味タグ体系は当面は固定せず、タグ体系の更新とタグづけを循環的に行う。参考までに、MSFA

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Frame ID	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14
2	F-to-F relations	targets F2	targets F3	elaborates F11; constitutes F4; presupposes F5	elaborates F13				constitutes F3	constitutes F3	targets F11	elaborates F12			presupposes F13
3	Frame Identifier	加圧 [+metaphoric]	押さえつけ [+metaphoric]	防御の際の相手の牽制	動作の継続	行動の先導	<節の接続>	行動の連続	集団行動?	攻撃	封印 [+metaphoric]	防御者による攻撃の無力化	大学ラグビーの試合	ラグビーの試合観戦	観戦内容の報告
4	*													観戦者	報告者
5	*[+anaphoric: target=明大]	加圧体	押えつける者	牽制者	継続者		共通項	行動体	集団行動体	攻撃対象	封印者	防御者	チーム1	試合内容	報告内容
6	京産大	加圧対象	押えつける相手	牽制相手	動作	全体	節1	行動1		攻撃体	手段1	手段1	チーム2	チーム2の主力選手	
7	の					MARKER									
8	ゲーム					先導体									
9	メーカー														
10	広瀬	加圧対象	押えつける相手	牽制相手											
11	に	MARKER	MARKER	MARKER											
12	圧力	GOV[+composite]	GOV[+composite]	GOV[+composite]											
13	を														
14	かけ														
15	続け				GOV										
16	、						EVOKER	EVOKER							
17	集団	目的	目的	目的	目的		節2	行動2	GOV		手段2	手段2			
18	パワー								GOV						
19	で										MARKER	MARKER			
20	京産大									攻撃体	攻撃者	攻撃者	チーム2		
21	の										MARKER	MARKER			
22	攻め									GOV	攻撃技	攻撃技			
23	手									手段: 攻撃技		GOV			
24	を										MARKER				
25	封じ										GOV				
26	た						EXT	EXT			EXT	EXT			
27	。														

図1 京大コーパスの一文 (S-ID : 950103083-006 KNP : 96/11/04 MOD : 96/12/03)

「京産大のゲームメーカー広瀬に圧力をかけ続け、集団パワーで京産大の攻め手を封じた。」のMSFA を使った意味タグづけ。個々の列がフレームに、形態素行とフレーム列との交点にある指定 (= フレーム要素) が「意味タグ」に相当する。

を使った意味タグづけの結果(断片)を図1に示す。

図1で使われる MSFA が実現する意味解析は、形態素への(適当な辞書から選んだ)語義の付与ではない。ヒトが(何らかの文脈の中で)文を読み/聞いたときに理解する内容の、従来にない細かい粒度での記述である。従来の意味記述にありがちな「(計算機上での安易な実装を見込んだ)形式化への(過度の)こだわり」を回避し、特に従来の語彙資源では記述されなかった連語的単位=超語彙的単位で成立する意味の記述に優れる[2][3]。

書きコトバ部門では今までに、(A)京都大学テキストコーパス(京大黒橋研究室が管理)から選んだ63文(3記事)(著作権保持者の毎日新聞から公開権を得ていない分としてもう1記事分ある)、(B)日英対訳データベース(NICT 内山将夫が開発・管理)から選んだ47文(5話)に意味タグづけを行い、その結果を次の二つのサイトで公開している(一部の結果は未整形のため未公開)：

Site 1 : <http://www.kotonoba.net/~focal/cgi-bin/hiki/hiki.cgi?FrontPage>

Site 2 : <http://www.kotonoba.net/~focal/cgi-bin/hiki2/hiki.cgi?FrontPage>

Site 2 で公開されているデータ A の公開分には、1,300 個前後、Site 1 で公開されているデータ B の公開分には 400 前後のフレームが認定されている。

2.2 MIFA を使った話しコトバの意味資源開発(野澤)

話しコトバ部門では、MSFA を使った書きコトバのタグづけと平行し、話しコトバを対象にして同様の枠組みで MIFA によるタグ付きコーパスの構築を行っている。これは、タグ付きコーパスの多様性に貢献する目的で、2006 年度より開始された。MSFA が書き言葉向けのタグづけ仕様であるのに対し、MIFA は話しコトバ向けのタグづけ仕様である[4]。また、タグづけ対象となる適切なコーパスが存在しないため、これも自前で作成している。MIFA の枠組みでは、これまでの言語処理に見られるような、話しコトバに特化したタグ体系は使用せず、書きコトバのコーパスと共有可能な、MSFA と互換性のあるタグ体系を使用している。

初年度は、国会議事録や「日本語話し言葉コーパス」に対する実験的なタグづけを通して、話し

言葉の取扱いにおいて必要な仕様を確立し、2007 年度からは、博報『ことばと文化・教育』研究助成を利用して、実際の商取引対話に対して、複数の作業者によるタグづけを行い、小規模ながらも、意味と相互行為の両面にわたる粒度の高いタグ付きコーパスの構築を目指している。MSFA/MIFA で初めてとらえられるような粒度の細かい相互作用の記述が十分な量あることが「ちゃんと動く」対話システムの開発に不可欠な要素であると考えている。

3 誤りを含む言語の解析に向けて—英語学習者データベース・NICT JLE Corpus—(和泉)

国際化・情報化の進展に伴い、言語障壁の克服の必要性が高まっている。外国語学習者がより効果的に言語を習得できるような教授法を考えるためには、教師や研究者が学習者の現状の言語運用能力と、今後どのような段階を経てより高度な言語運用能力を身に付けていくのか、学習者の発達段階を全体的に把握しておくことが不可欠である。そのためには、学習者言語の普遍性・多様性の両方を測るに足るだけの規模の学習者言語の実例データを収集し、信頼のおける解析結果に基づいた議論をすることが必須である。また、コンピュータ技術の発展に伴い、コンピュータを介したコミュニケーションや、CALL (Computer-Aided Language Learning)・e-learning のような新しい学習環境の導入が徐々に盛んになってきており、外国語学習者とコンピュータ間のコミュニケーションの支援が必要である。

従来、自然言語処理においては、基本的に文法的・意味的に正しい言語の機械処理を対象に研究が行われてきた。しかし、CALL や e-learning のような人間とコンピュータとの新しい環境が作られている状況で、非母語話者が話す言語に含まれる誤りをコンピュータでうまく解析する枠組みの構築に対する必要性が高まっている。学習者言語のコーパスは、そのために必要な誤りの実例を含む学習者言語の振る舞いのサンプルデータであり、英語教育や外国語習得研究の発展、そして学習者言語の機械処理技術の開発のために有効な素材となり得る。本章では、日本語を母語とする英語

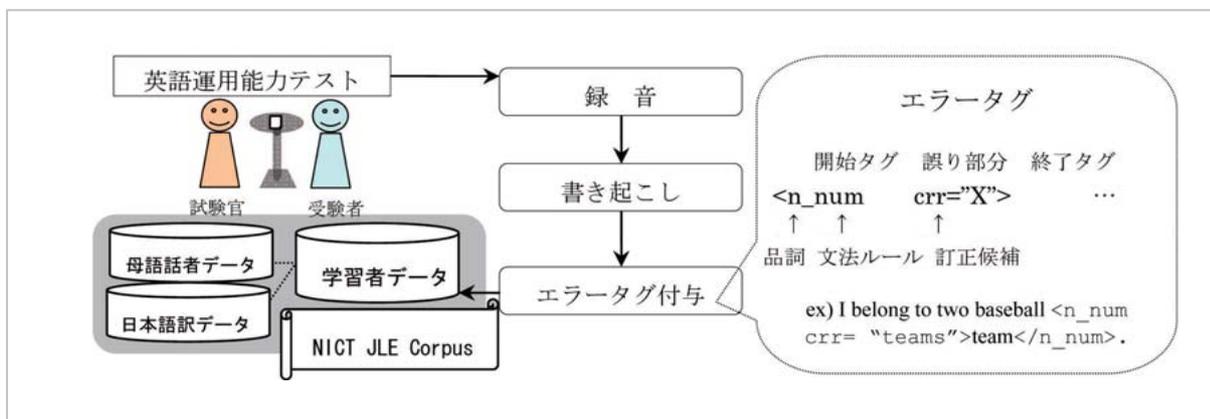


図2 NICT JLE Corpus の構成

学習者の英語発話のデータベースである、“NICT JLE (Japanese Learner English) Corpus” [5] を用いた研究について述べる。

3.1 NICT JLE Corpus の概要

NICT JLE Corpus (図 2) は、日本語を母語とする英語学習者 1,281 名分(延べ 200 万語)の英語インタビューテスト(1 名につき 15 分間)の音声を書き起こしたテキストデータで構成されている。各データには、発話者の熟達度レベルが 9 段階で示されている。また、本コーパスのデータの一部(167 名分)には、文法的・語彙的誤りを対象とした 47 種類のエラータグが人手によって付与されている。学習者の誤りを質的・量的に分析することは、学習者の習得段階をモデル化するための有効な手段の一つである。ただし、エラータグにおいて扱われている誤りの種類は、文法、語彙といった比較的形式的なものであるため、学習者言語を更に広い視野で観察することを可能にするために、二つの補助的コーパスを作成した。一つ目は、学習者の英語を話者の意図をできる限り推測し正しい日本語に訳した日本語訳コーパスである。日本語訳データとエラータグ付き学習者英語データを合わせて観察することにより、外国語習得における母語の干渉度合いを測ることが主なねらいである。二つ目は、英語母語話者による発話コーパスである。母語話者の発話と学習者の発話を比べることにより、使用語彙頻度や話の進め方の違いなど、エラータグだけでは観察できない知見を見いだすことができると考える。

3.2 学習者言語の誤り分析と学習支援システム開発への応用

NICT JLE Corpus を利用して、学習者言語の誤り分析を中心とした学習者言語の記述や学習支援システムの開発を行っている。分析・開発の際の着眼点は、第二言語習得研究において古くから行われてきた誤り分析(Error Analysis)の手順(図 3)からヒントを得ている。

まず、「誤りの確認」は、学習者の発話のうちどの単語・フレーズ、文構造、語順が誤っているか、という誤りが起こっている箇所を特定することである。「誤りの分類」は、その誤りの言語範ちゅう(形態素・統語・語彙など)が正用法とどう異なるかによって、誤りを分類することである。次に、「誤りの説明」とは、何故その誤りが引き起こされたのか、誤りの原因を特定することである。最後に、誤りの評価は、誤りを含む文の分かりやすさを評価することである。つまり、様々な誤りをコミュニケーションの大きな妨げになるものとそれほど影響を及ぼさないものに分類し、誤りの重み付けをすることである。

誤りの確認と分類に相当するタスクとして、前述の 47 種類のエラータグに基づき、学習者の誤りの特徴に関する幾つかの分析を行った。冠詞の誤りやイディオムの誤りなど個別の言語項目に焦点を当てた、それらの出現パターン(出現形態や出現文脈)の分析[6]や、主要文法形態素の習得順序の抽出(図 4)[7]などを行った。これらの分析から得られた知見を元に、機械学習を用いて誤りの自動検出実験を行い、自動誤り検出システム Eden (Error Detection in English) を開発した[8]。

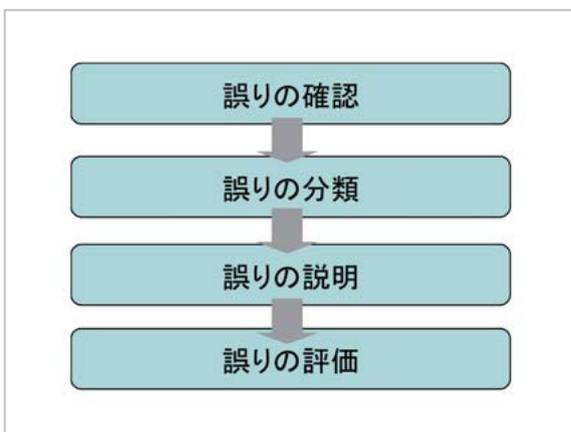


図3 誤り分析の手順

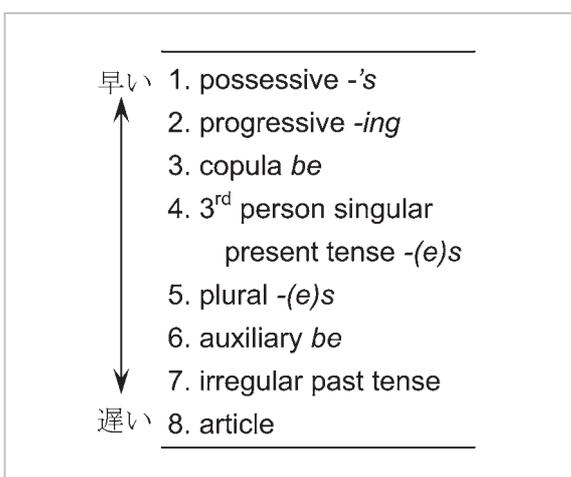


図4 日本人英語学習者の主要文法形態素習得順序

ここまでの分析・実験では、誤りがあるかないか、といった学習者言語の正確性に焦点を当てた。外国語学習において、その目的によっては言語の正確さを追求することは重要となるが、近年主流となっているコミュニケーション優先の言語学習を行う場合、まずは絶対に誤ってはいけない項目と、必ずしも正確さが要求されない項目は何かを知っておくことは有益である。そこで、本コーパス利用の第二フェーズとして、正確さよりも「通じやすさ」に焦点を当て、学習者言語をよりコミュニケーション視点から分析することに焦点を当てた。これは、図3で示した誤り分析の手順のうち、「誤りの評価」に相当するタスクと考える。まず NICT JLE Corpus のデータに対して、各文の通じやすさのレベルを三段階 (intelligible, unclear, unintelligible) で英語母語話者に判定してもらった。その結果、各文に含まれる誤りの種

類・頻度が、各文の通じやすさのレベルに違いを及ぼす影響を持つことが分かった[9]。この誤りの種類と通じやすさの関係を利用して、機械学習に基づく通じやすさレベルの自動判定実験も行った[10]。intelligible かそうでない (unclear もしくは unintelligible) かは、90 %以上の精度で判定できたが、文脈レベルでの情報を必要とする unclear と unintelligible の区別は非常に困難であった。

本節では、NICT が持つ言語資源のうち、従来の言語処理では対象とされなかった、誤りを含む文の解析に貢献し得る学習者コーパス・NICT JLE Corpus の概要と、それに基づく学習者言語の分析と機械処理技術の開発について述べた。今後は、基礎研究的な検証を実用レベルの技術へと応用すべく開発を続けていく予定である。

4 NICT 日中機械翻訳ための言語資源の構築 (張)

機械翻訳の研究又はシステムの開発の一環として、NICT では、日中コーパスと日中翻訳辞書の構築を行っている。

4.1 日中对訳コーパス [11]

対訳コーパスは、異なる言語の対を収集したもので、文章、段落あるいは文を単位として構成される。対訳コーパスは源言語と目的言語との間の異なるレベルでの対応関係を網羅することができるので、例えば機械翻訳の研究や機械翻訳システムの開発において翻訳知識を抽出する際などに必須となる言語資源である。NICT では、アジア言語を含む多言語コーパスを構築するプロジェクトを数年前にスタートした。現在は、日本語と英語、日本語と中国語の対訳コーパスを統一した仕様で構築中である。ここでは、日中对訳コーパスを紹介する。

(1) 日本語データ

日本語データは 1995 年の毎日新聞の記事から抜粋された約 4 万文からなる。日本語データに対しては、日本語話し言葉コーパスの仕様に基づいて、単語分割、品詞タグ付け作業、そして構文構造タグ付け作業を行う。

(2) 中国語データ

中国語データは上記日本語データを中国語に翻

訳したものである。翻訳はプロの翻訳者によって行われた。翻訳基準は以下のものになる。

- (a) 一文の日本語を単位として訳す。
- (b) 構造が原文に近い訳文を優先する。
- (c) 意味が通じるように、必要があれば前の文から情報を補う。特に、日本語には、主語がよく省略されるが、中国語には主語が必要なので、訳文には主語を補う。
- (d) 流暢性、読みやすさを考慮し、必要に応じて語順の入れ替え、カンマ「,」の挿入などを行う。訳文の質を確保するために、翻訳者による精密化と中国語のネイティブによる精密化を行う。

(3) 中国語文の形態素解析と人手修正作業

中国語訳文に対して単語分割及び品詞タグ付け作業を行う。中国語の品詞体系においては 39 個の品詞が定義されている。まず、解析ツールを用いて単語の分割と品詞タグ付けを行う。そして、得られた結果に対し人手により修正を加える。人手修正作業の効率を高めるために、補助ツールを開発した。補助ツールは一般の編集機能以外、以下の便利な機能を持っている。

- (a) 「現代漢語文法情報辞書」への単語検索ができる。
- (b) チェック済みコーパスに対し単語検索ができ、検索結果をその単語の左右の文脈によってソートすることができる。
- (c) 直前に変更された単語と同じ属性を持つすべての単語に対し、同じ変更を一括で行うことができる。この機能は単語の分割と品詞タグ付けにおいて一貫性を確保するときにより便利である。

(4) 日中対訳コーパスにおける単語アライメント^[12]

大規模な日中対訳コーパスに対して、単語・句レベルでのアライメントを行い、翻訳知識としての対訳関係を抽出することが必要である。この研究では、統計情報に基づく手法と語彙知識を利用する手法のそれぞれの長所を取り入れられるような統合的な単語アライメント手法を提案し、評価実験を行う。

- (a) 語彙知識を利用する手法は二段階の手順、語彙情報に基づくアライメントと位置情報に基づくアライメントからなる。日本語文

の任意の形態素と中国語訳文の任意の単語の対に対し、その間に対訳関係がある可能性を推定して、その中からもっともらしいものを決定すればよい。可能性の推定には三種類の語彙情報、翻訳辞書、漢字の字形情報及び漢字の簡体字と繁体字の関係を利用した。語彙情報で対応付けできていない日本語形態素と中国語単語に対し、位置情報に基づいてアライメントをする。一般には、原文において同じ構文成分に属する単語が訳されたとき、対訳関係にある訳語も訳文において同じ構文成分に属することがよく観察される。この現象により、既に得られたアライメント情報を利用することによって行うことができる。日本語文の形態素 j と中国語訳文の単語 c の間に対訳関係がある可能性を推定するには、 $j(c)$ の左側に一番近いアライメントと $j(c)$ の右側に一番近いアライメントの四つのアライメントを利用する。 j と c との間のアライメント候補と四つのアライメントのそれぞれのねじれ及び離れ具合を計ることにより、可能性を推定する。

- (b) 統計情報に基づく手法としては、既存の統計手法 GIZA++ ツールを利用した。
- (c) 統合的な手法

語彙情報と統計情報のそれぞれの長所を生かす統合的な手法を試みた。具体的には、(a) の手法、GIZA++ の $C \rightarrow J$ と $J \rightarrow C$ の多数決の結果を採用することにした。この手法を評価するために、日中対訳コーパスから 1,127 文対を抽出し、人手で単語の対応付けを付与した。評価の結果、再現率が 63% に、正解率が 79% に向上した。したがって、語彙情報と統計情報を同時に利用することが有効であることが分かった。自動的に得られた結果に対し、人手による修正を加える。

4.2 EDR 日英辞書の中国語への拡張^[13]

対訳辞書は機械翻訳や言語横断検索において欠かせない言語資源である。この研究では、EDR 日英辞書^[14]を中国語へ拡張し、日中の基本対訳辞書の構築を行う。それにより、日英中三言語の

対訳辞書が得られる。EDR 日英辞書の各レコードに対し、以下のような情報を付与する。

(a) レコードの<意味情報>に基づいた中国語訳語。(b) 中国語訳語ごとの品詞・構文情報。(c) 中国語訳語ごとのレジスター情報。(d) <日本語概念見出し>への中国語訳。(e) <日本語概念説明>への中国語訳。(f) 中国語訳語のバリエーション情報。

レジスター情報に以下のような情報がある。

(a) 中国語訳語の種類：「逐語訳」、「説明文」、「音訳」など。(b) スタイル：「古語」、「口語体」、「文語体」、「俗語」。(c) フォーマル：「尊敬」、「謙譲」、「軽蔑」。

中国語訳語に品詞・構文情報を付与するには、まずその訳語が単語か句かを判別する。次に、単語あるいは句の分類に基づき品詞・構文情報を付与する。単語の品詞情報は 18 種類とし、句の構文情報は 11 種類とした。

参考文献

- 1 C. R. Johnson and C. J. Fillmore, "The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure", In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000), pp.56-62, 2000.
- 2 K. Kuroda, M. Utiyama, and H. Isahara, "Getting deeper semantics than Berkeley FrameNet with MSFA", In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06), P26-EW, 2006.
- 3 黒田 航, 井佐原均, "均意味フレームを用いた知識構造の言語への効果的な結びつけ", 信学技報 104 (416), pp.65-70, 2004.
- 4 野澤 元, 黒田 航, 井佐原均, "対話テキストにおける意味と談話：複層意味フレーム分析による試み", NLC2006-28 (2006-10), 信学技報, 106 (299), pp.33-38, 2006.
- 5 和泉絵美, 内元清貴, 井佐原均, "日本人 1200 人の英語スピーキングコーパス", アルク, 2004.
- 6 和泉絵美, 齋賀豊美, Thepchai Supnithi, 内元清貴, 井佐原均, "エラータグ付き日本人英語学習者発話コーパスを用いた学習者の冠詞習得傾向の分析", 言語処理学会第 9 回年次大会発表論文集. pp.19-22, 2003.
- 7 和泉絵美, 内元清貴, 井佐原均, "エラータグ付き学習者コーパスを用いた日本人英語学習者の主要文法形態素の習得順序に関する分析", 言語処理学会論文誌『自然言語処理』特集号『コーパス言語学・CALL と言語処理』, pp.212-225, 2005.
- 8 E. Izumi, T. Saiga, K. Uchimoto, T. Supnithi, and H. Isahara, "Automatic error detection in the Japanese Learners' English spoken data", In Companion Volume of the Proceedings of the Association of Computational Linguistics (ACL) 03. Japan, pp.145-148, 2003.
- 9 和泉絵美, 内元清貴, 井佐原均, "日本人英語の通じやすさに関する研究", 『言語処理学会第 12 回年次大会発表論文集』, 2006.

中国語化の作業は、平成 17 年度から始めた。中国国内の有力な専門翻訳者と豊富な中国語処理資源を活用するようにしている。現在頻度の高い日本語単語 (JUMAN の日本語単語リストと重なる部分) をはじめ、約 14 万個のレコードが完成された。平成 19 年度まで全部完成して公開する予定である。

5 むすび

インターネットの普及に従って自然言語処理の技術はますます重要になっていく。自然言語処理の応用は機械翻訳、言語習得支援などがある。この論文で紹介したような深いレベルの情報タグ付きコーパスは、言語処理の応用の性能を高めるための、深層的な自然言語処理に貢献できるものと期待している。

- 10 E. Izumi, K. Uchimoto, and H. Isahara, "Measuring intelligibility of Japanese learner English, T. Salakoski, F. Ginter, S. Pyysalo, & T. Pahikkala, (Eds)". Advances in Natural Language Processing, 5th International Conference on Natural Language Processing (FinTAL), Lecture Note in Artificial Intelligence, Springer, Berlin, pp.476-487, 2006.
- 11 Y. Zhang, K. Uchimoto, Q. Ma, and H. Isahara, "Building an Annotated Japanese-Chinese Parallel Corpus-A Part of NICT Multi lingual Corpora", In the Tenth Machine Translation Summit Proceedings, pp.71-78, 2005.
- 12 Y. Zhang, Q. Liu, Q. Ma, and H. Isahara, "A Multi-aligner for Japanese-Chinese Parallel Corpora", In the Tenth Machine Translation Summit Proceedings, pp.133-140, 2005.
- 13 Y. Zhang, Q. Ma, and H. Isahara, "Automatic Construction of Japanese-Chinese Translation Dictionary Using English as Intermediary", Journal of Natural Language Processing, 12(2), pp.63-85, 2005.
- 14 NICT (National Institute of Information and Communications Technology), "EDR Electronic Dictionary Version 2.0 Technical Guide", 2002.



張 玉潔

知識創成コミュニケーション研究センター自然言語グループ専門研究員(旧情報通信部門けいはんな情報通信融合研究センター自然言語グループ長期専攻研究員) 博士(工学)
自然言語処理



黒田 航

知識創成コミュニケーション研究センター自然言語グループ専攻研究員(旧情報通信部門けいはんな情報通信融合研究センター自然言語グループ長期専攻研究員) 博士(人間・環境学)
言語学・認知科学



和泉 絵美

知識創成コミュニケーション研究センター自然言語グループ専攻研究員(旧情報通信部門けいはんな情報通信融合研究センター自然言語グループ短期専攻研究員) 博士(学術)
コーパス言語学・第二言語習得



野澤 元

知識創成コミュニケーション研究センター自然言語グループ専攻研究員(旧情報通信部門けいはんな情報通信融合研究センター自然言語グループ短期専攻研究員)
認知言語学・語用論