様式 C-7-2

自己評価報告書

平成　２２　年　５　月　７　日現在

研究種目：基盤研究（B）
研究期間：2007～2010
課題番号：　19300047
研究課題名（和文）　LCTL を含む多言語平行マルチメディア資源の構築と構造化方式の研究

研究課題名（英文）　Construction of multimedia parallel multi-language resources including LCTL and Research of Structured Language Schemes
研究代表者
堀　一成（HORI KAZUNARI）
大阪大学・大学教育実践センター・准教授
研究者番号：80270346

研究代表者の専門分野：自然言語処理、多言語資源構築、数理工学
科研費の分科・細目：情報学・知能情報学
キーワード：外国語、多言語処理、コンテンツ・アーカイブ、言語資源、LCTL

１．研究計画の概要

（１）多言語対応構造化方式の研究

産業技術総合研究所の橋田氏が提案されているGDA（大域文書修飾）を基盤とし、多言語平行の言語資源を構築するに際して、さらにふさわしい構造化方式について研究し、提案する。

（２）諸言語の文字・音声データ集積

構造化された基礎語彙・会話データに関する文字・音声データの集積を行う。全期間で１０言語分のデータ集積を計画している。その成果を言語工学系分野に多言語資源として提供し、周知をはかる。

（３）　構築資源の教育への応用

構築した資源をデータベース化し、言語教育の教材として応用をはかる。

（４）携帯端末での多言語情報発信応用

携帯端末をターゲットとして、構築言語資源を活用するシステム技術の研究を行う。

２．研究の進捗状況

（１）多言語対応構造化方式の研究

１９年度は、独自に選定した「言語学的に興味深い文 100」（オリジナルは日本語）に，統語情報を示すタグをつける作業を進め，完成させた。また統語情報タグをつける作業のための FLASH アプリケーションの開発に着手した。

２０年度も作業を進め、100 文タグ付加作業は、日本語・中国語が完成し、朝鮮語の翻訳作業も進行させ、論文に発表した。オントロジー情報も付与すべきであるとの結論に至り、作業を補助するための FLASH ソフトウェアを開発した。

２１年度は、これまで開発した機能を統一し、オントロジー情報を XML データの属性として付与できるよう FLASH アプリケーションを改良した。このツールを用いて言語資源構築作業をするうち、GDA の多言語対応に対する不備な点が明らかになり、その改良となる方式をまとめ提案すべく作業進展中である。

（２）諸言語の文字・音声データ集積

１９年度においては、ペルシア語の会話データについて、試験的に分野ごとによる GDA 化 XML データ化の作業を行った。文単位で ID 番号を付与し、言語名を識別するためのデータを付与した。

２０年度においては、ロシア語の言語資源構築について、1000 以上のフレーズを含む会話集のロシア語訳を行った。GDA 化の作業を進めていたペルシア語の会話データについては、データとしての質を一層高めた。特に GDA 構造化情報はレベル３（詳細な統語構造）までの付与に進展した。その作業において人称語尾の変化を動詞のタグ内に埋め込む必要があるなど、改良すべき問題点が発見された。その成果を「イラン研究」誌に論文として発表した。

２１年度、ペルシア語資源については、基本会話文 1000 文の GDA による構造化の作業を補足的に行った。これに加え、日本語の基本会話文 1000 文の GDA による構造化の作業も行った。補足的な作業は今後もあり得るが、ほぼ作業は終了した。ロシア語の言語資源構築については、会話集の録音を完了し、音声資料の編集作業を開始した。　トルコ語資源については、ペルシア語データを参考に、基

本的な会話文100文程度を選び、同内容のトルコ語会話文データとそのGDAタグ付け作業を進めた。

（３）　構築資源の教育への応用

２０年度に、分担者（上原）がロシア語資源を教材として応用するための試みを行い、その成果を論文として発表している。

（４）携帯端末での多言語情報発信応用

１９年度は、携帯端末への応用としてWindows Mobile端末への資源搭載作業を行ったがフォントの扱いに問題があることが判明した。そのため２０年度、２１年度とWindows Vista 搭載でフィールドでの利用を想定した堅牢PCを購入し、言語資源搭載と、利用のためのソフトウェア開発を行った。成果を非公開型科学技術交換会で発表した。

３．現在までの達成度

②　おおむね順調に進展している。

（理由）

エンコーディング方式の研究においては、が新方式の内容をほぼ固め、提案に向け準備しているところである。オントロジー情報や統語情報をエンコーディングする作業のためのFLASHアプリケーションの開発は、当初計画になかった新しい成果であり、計画以上といえる。資源の教育応用についてはロシア語資源の活用事例を論文で報告している。言語資源を堅牢 PC に搭載し、フィールドや教育の場で提供するための技術研究・ソフトウェア開発も一定の成果を収め、科学技術情報交換会で発表することが出来ている。一方、構造化済み言語資源は、３年目終了時点で７言語完成している予定であるが、現時点で５言語であり、完成に至っていないものもあり、やや遅れている状況である。以上のように計画以上のところと、遅れのあるところを勘案し、表記の評価になると考えた。

４．今後の研究の推進方策

各研究の完成に向けて活動を進める方針は以下のとおりである。

（１）　多言語対応構造化方式の研究

個別言語データによる検証から得られる諸言語固有の問題と、言語処理上必要なタグの追加、修正の研究を行う。特に分担者（山崎）がエンコーディング作業を試行したことにより、GDAに代わる新たなタグセットを提案できる段階に至っている。成果をまとめ、提案とする。また統合情報・オントロジー情報などの言語学的情報を付与するための作業を容易にするソフトウェアを平成20年度から分担者（鈴木）を中心に開発中であり、完成度を上げる作業を行う。

（２）諸言語の文字・音声データ集積

最終年度のデータ集積作業は、ロシア語、トルコ語、英語、アラビア語の完成を目標に作業を進める。特に、前項で検討する構造化手法を基盤にデータのタグ付け作業を実施する。分担者（竹原・上原）および連携研究者と協力し作業を進める。分担者（小島・竹原）は、資源を利用した翻訳技術教育のための簡易アプリケーションの開発も行う。

（３）携帯端末での多言語情報発信応用

連携研究者と、在住外国人のための行政情報提供に関して意見交換を行い、堅牢ノートPCに言語資源を搭載し、それを教育や社会貢献に活用してもらえるよう開発を行う。

５．　代表的な研究成果

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　３ 件）

① 竹原新、GDA によるペルシア語文の構造化、イラン研究、第5号、pp.159-177、2009年、査読有

② 上原順一、XML を用いたロシア語の語形成電子教材の可能性について、大阪大学　世界言語研究センター論集、第一号、pp.63-73、2009年、査読有

③ 山崎直樹、多言語平行コーパスのための「言語学的におもしろい 100 の文」、外国語教育研究、pp.1-15、2009年、査読有

〔学会発表〕（計　４ 件）

① 鈴木慎吾、山崎直樹、堀一成、多言語資源作成のための統語属性付与支援FLASHアプリケーションの開発、言語処理学会第 16 回年次大会、2010年3月10日発表、東京大学本郷キャンパス

② 堀一成、萬宮健策、山根聡、平松初珠、石島悌、災害救援者のための多言語データ蓄積と関連する携帯 PC 向けアプリケーションソフトウェアの開発、第4回非公開型科学技術情報交換会、2009年12月16日発表、大阪国際会議場

③ 鈴木慎吾、山崎直樹、堀一成、テキストコーパスにオントロジー的知識を付与するためのFLASHアプリケーションの開発、言語処理学会第15回年次大会、2009年3月3日発表、鳥取大学　鳥取キャンパス

④ 鈴木慎吾、山崎直樹、堀一成、多言語資源作成のための文構造タグ付加支援FLASHアプリケーションの開発、言語処理学会第 14 回年次大会、2008年3月18日発表、東京大学　駒場キャンパス

〔その他〕

研究内容紹介ホームページ

http://mldb.cep.osaka-u.ac.jp