様式 C-19

# 科学研究費補助金研究成果報告書

平成 23 年 5 月 20 日現在

機関番号：14401
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

研究成果の概要（和文）：
本研究は、大阪大学の多言語資源研究グループを中心とし、文字・音声情報を含む多言語間比較可能なパラレルコーパス（平行言語資源）の構築と、その構築作業をサポートする XML 情報付加 Web ソフトウェアの開発を行ったものである。作成した言語資源の内訳は、5000 語の基本単語集（7 言語）、旅行時の会話を中心とする 1000 文会話文集（12 言語）、前記会話集に文構造を示す XML タグを付与した構造化文集（5 言語）である。文構造を表す XML タグのセットは産業技術総合研究所で開発された GDA: Global Document Annotation を採用した。言語資源作成作業の補助となる XML 情報付加ソフトウェアは、Flash システムを用いた Web プログラムとなっており、GUI 操作により直感的に作業者が文の木構造を創り上げることができ、同時に XML データも自動生成する仕様となっている。

研究成果の概要（英文）：
The Multi-language Resources Research Group of Osaka University has been working on multi-language parallel corpora and XML annotation tools. The contents of the corpora include 5000-word lists (seven languages), 1000 plaintext sentences (12 languages), and XML-formatted sentences (five languages) that contain syntactic information. Each corpus has words and sentences which are organized in a tabular format. The type of XML format used for these corpora is Global Document Annotation (GDA), which allows computers to automatically recognize the semantic and pragmatic structures of texts. Three XML annotation tools are created especially for tagging words and sentences in these corpora, in forms of FLASH applications so that they can be used on Web browsers. The tools' GUI has a function to draw parse trees, which helps annotators who are not familiar with XML data construction. These corpora can be used as fundamental data for contrastive linguistics and comparative linguistics, and also be used as training data to verify the validity of statistical analysis in natural language processing researches.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合 計 |
|---|---|---|---|
| ２００７年度 | 4,900,000 | 1,470,000 | 6,370,000 |
| ２００８年度 | 3,200,000 | 960,000 | 4,160,000 |
| ２００９年度 | 2,900,000 | 870,000 | 3,770,000 |
| ２０１０年度 | 3,400,000 | 1,020,000 | 4,420,000 |
| 年度 | | | |
| 総 計 | 14,400,000 | 4,320,000 | 18,720,000 |

研究分野：言語資源研究
科研費の分科・細目：情報学・知能情報学
キーワード：自然言語処理、多言語資源、会話文データベース、語彙データベース

１．研究開始当初の背景

研究開始当初は、IT技術の飛躍的発展や携帯端末の高機能化を背景として、ユビキタス情報化社会に向けた基礎・応用研究が急速に進展していた時期であった。日本国内で現在も進行している多言語化した社会の現実からも、多言語間の情報バリア・フリーを求める社会状況は、切実であった。しかし、工学系分野の貢献によりこれらを実現するハード的技術は整いつつあったとはいえ、多言語間の情報処理技術（「多言語機械処理」）における発展を伴わないならば、国内外における多言語社会のための、真の情報バリア・フリーを実現することは不可能であると考えられた。

２．研究の目的

前記のような背景から、言語工学系分野の諸研究者の支援を得て連携しつつ、大阪大学の外国語研究者が持つ知見を結集して、多言語機械処理に不可欠な言語面に関わる基礎技術の発展に貢献し、かつ、現行の技術で対応可能な応用研究を行おうとするのが、本研究の目的である。

LCTL (Less Commonly Taught Languages) とは、英語・日本語・中国語などの主要言語に比較して、教育や研究の対象となる機会の少ない言語である。しかし、昨今南アジア・中東地域を含む各国の国際社会における重要性は増し、自然言語処理においてもその対象とすべきであるが、基盤となる言語資源の蓄積は十分であるとはいえない。

本研究では、自然言語処理研究の重要な基礎データとなりうる多言語を対照比較できるよう、特に構文情報など言語学的研究に有用な情報をXMLタグとして付与した平行言語資源の作成を主目的とする。また、そのような作成を担当する作業者の労力を軽減するための作業支援ソフトウェアも開発する。あわせて、作成したデータを言語教育に利用する取り組みも行う。

３．研究の方法

（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（Global Document Annotation: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）言語資源構築支援ソフトウェアの開発

研究した構造化方式に従って、会話文に構文情報や意味関係のオントロジー情報を付与する作業を補助するためのソフトウェアを開発する。作業環境の自由度を上げるため、固定のPCへのインストールを必要としないWebアプリケーションでの実現を目指す。

（３）諸言語の文字・音声データ集積

5000語の基本単語集、旅行時の会話を中心とする1000文会話文集、XMLタグにより構造化された会話文集データに関する文字データの集積を行う。可能なかぎり対応する音声データの集積も行う。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。

（４） 構築資源の教育への応用

構築した資源を、分担者が自らの担当する講義中で利用し、言語教育の教材として応用する。 自然言語処理研究以外の利用法から新たなタグ情報案の提案やソフトウェア開発のアイデアを得る。

（５）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、構築言語資源を広範な場所で活用するシステム技術の研究を行う。

４．研究成果

（１）多言語対応構造化方式の研究

平成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。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。次項で説明する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）言語資源構築支援ソフトウェアの開発

まず、平成19年度に、文の構文を表す統語情報タグをつける作業のためのWebアプリケーションの開発に着手した。GUIでの操作が可能であり、木構造を簡単に表示できること、操作状況をアニメーションにより把握できることが必要と考え、開発システムはAdobe Flashで行うこととした。

20年度において、単語間の意味つながりの関係を表すオントロジー情報も付与すべきであるとの結論に至り、統語情報付与アプリケーションと同じくFlashプラットフォームにおけるソフトウェアを推進した。 21年度から22年度に掛けて、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アプリケーションを改良した。この成果は第9回情報科学技術フォーラムに査読有論文として投稿するとともに、発表を行った。

以下に完成したWebアプリケーションの操作時画面キャプチャー画像を掲載する。

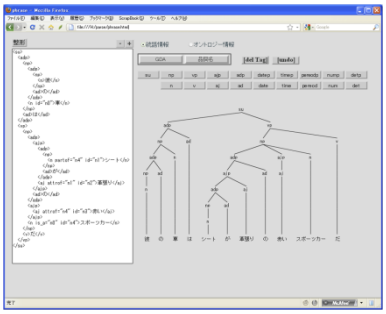

**図 1　構文情報付与作業時のアプリケーション実行画面例**

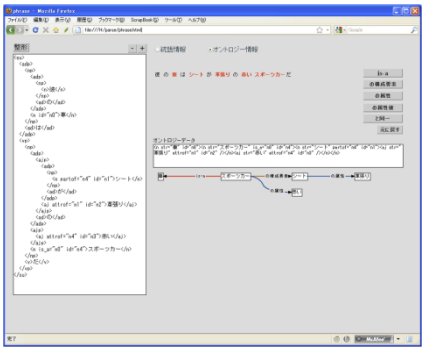

**図 2　オントロジー（意味関係情報）付与作業時のアプリケーション実行画面例**

（３）諸言語の文字・音声データ集積

平成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。

22年度は、英語のGDAに基づくXML構造化会話文集を完成させた。また、前年度から着手していたトルコ語およびロシア語会話文のXML構造化作業を、挨拶文などを中心とする基本会話部分（100文強）について進展させた。

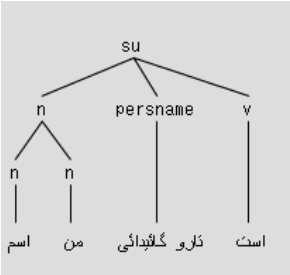


**図 3　GDA タグ付与作業済ペルシア語資源の一文を作業支援ソフトウェアで木構造表示した例**

```
<su syn="f">
  <n syn="b" opr="aen">
  <n>اسم </n>
  <n>من </n>
  </n>
  <persname>تارو گائیدائی</persname>
  <v>است </v>
.</su>
```

枠内のデータは、図3に示した文例のXMLデータを例示したものである。

本研究終了時点での蓄積言語資源量をまとめると、以下のとおりとなる。

基本5000単語（7言語:日・英・ヒンディー・ペルシア・アラビア・中国・朝鮮）

会話文集1000文（12言語:日・英・アラビア・スペイン・トルコ・ヒンディー・ペルシア・モンゴル・中国・朝鮮・ベトナム・タイ）

XMLタグ付与済会話文集 5言語

（1000文完成:日・英・ペルシア、100文完成:トルコ・ロシア）

（４） 構築資源の教育への応用

平成20年度から22年度にかけ分担者（上原）が作成した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。

**図 4　構築ロシア語資源を教材として講義中に活用している様子**

分担者（山崎）は22年度には 資源作成用Webソフトウェアを初心者向け中国語教育における読解補助教材・構文と文法事項の説明教材として活用している。この成果を22年度情報教育研究集会で発表した。

（５）携帯端末での多言語情報発信応用

平成 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 20～22 年度に 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。本ソフトウェアの特徴は、文字情報だけでなく、内容が対応する画像や動画も表示できる機能を有することと、XML 化した言語資源の処理に対応していることである。また、単独の Windows アプリケーションになっており、複雑なインストール操作をすることなく使い

始めることができることも利点である。


**図 5　携帯端末用アプリケーションの表示画面例（ウルドゥー語）**

本成果は大阪府産技研の連携研究者が担当し、非公開型科学技術交換会などで発表した。

５．主な発表論文等
（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１１件）

① 上原順一、MeCabで利用可能なロシア語辞書について、言語文化研究、第37号、pp. 315-322、2011年、査読有

② 平松初珠、石島悌、災害救援者用多言語データを活用したシステム構築、商工振興, No. 737、pp. 15-16、2011 年、査読有

③ 鈴木慎吾、山崎直樹、堀一成、多言語資源作成のための統語・オントロジー情報を付与するアプリケーションの開発、第９回情報科学技術フォーラム講演論文集　第４分冊、pp. 119-122、2010 年、査読有

④ 平松初珠、石島悌、災害救援者用多言語データを活用したシステム構築、　大阪府立産業技術総合研究所テクニカルシート, No. 09008, 2010 年、査読無

⑤ 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

⑥ 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

⑦ 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、第 17 号、pp. 111-125、2009 年、査読無

⑧ 山崎直樹、送り仮名・返り点付き漢文資料からどのような言語学的情報が得られるか、漢字文献情報処理研究、第 9 号、pp. 20-28、2008 年、査読有

⑨ 竹原新、イラン民話の抽象的様式、イラン研究、第4号、pp. 109-123、2008年、査読有

⑩ 山崎直樹、訓点付き漢文の返り点から統語情報を導出しＸＭＬで構造化する試み、漢字文献情報処理研究、第 8 号、pp. 73-82、2007年、査読有

⑪ 高橋明、小島一秀、岩成英一、これまでのeラーニングと問題集システム、コンピュータ＆エデュケーション、Vol. 23、pp. 30-35、2007年、査読有

〔学会発表〕（計１０件）

① 藤家洋昭、Reyihan Pataer、ウイグル語における再帰代名詞の人称、言語処理学会第17回年次大会、2011年3月9日発表、豊橋技術科学大学

② 堀一成、竹原新、上原順一、小島一秀、藤家洋昭、萬宮健策、GDAに基づく統語情報付与XML化多言語並行資源の構築、言語処理学会第17回年次大会、2011年3月8日発表、豊橋技術科学大学

③ 鈴木慎吾、山崎直樹、堀一成、多言語資源のための統語・オントロジー情報付与アプリケーションの開発と外国語教育での活用、平成22年度　情報教育研究集会、2010年12月11日発表、京都府民総合交流プラザ　京都テルサ

④ 鈴木慎吾、山崎直樹、堀一成、多言語資源作成のための統語・オントロジー情報を付与するアプリケーションの開発、第９回情報科学技術フォーラム、2010年9月9日発表、九州

大学 伊都キャンパス

⑤ 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

⑥ 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

⑦ 山崎直樹、多言語平行コーパスのための言語横断的な構造記述、2009中日理論言語学国際フォーラム、2009年7月26日発表、同志社大学 寒梅館

⑧ 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

⑨ 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

⑩ 高橋明、小島一秀、岩成英一、異文化障壁を乗越える対話と交渉能力の育成－実践的eラーニング言語教育プログラムの展開－、平成19年度大学教育改革支援プログラム合同フォーラム、 2008年2月9，10日発表、パシフィコ横浜　会議センター

〔その他〕

報道関連情報

日刊工業新聞　2010 年 2 月 11 日付紙面第 15 面　テクノ編集局 No.41「多言語処理」の記事内で、本研究の成果が紹介された。

ホームページ等

一連の研究成果は以下の Web ページ上で公開している。

「大阪大学　多言語資源研究グループ Web ページ」

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

６．研究組織

(1)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

(2)研究分担者

竹原　新（TAKEHARA SHIN）
大阪大学・世界言語研究センター・准教授
研究者番号：20324874

山崎　直樹（YAMAZAKI NAOKI）
関西大学・外国語学部・教授
研究者番号：30230402

小島　一秀（KOJIMA KAZUHIDE）
大阪大学・サイバーメディアセンター・講師
研究者番号：60372637

上原　順一（UEHARA JUNICHI）
大阪大学・言語文化研究科・准教授
研究者番号：30252737

鈴木　慎吾（SUZUKI SHINGO）
京都産業大学・外国語学部・助教
研究者番号：20513360

(3)連携研究者

石島　悌（ISHIJIMA DAI）
大阪府立産業技術総合研究所・情報電子部・主任研究員
研究者番号：80359398

高階　美行（TAKASHINA YOSHIYUKI）
大阪大学・世界言語研究センター・教授
研究者番号：70144540

藤家　洋昭（HUZIIE HIROAKI）
大阪大学・世界言語研究センター・准教授
研究者番号：90283837

谷村　緑（TANIMURA MIDORI）
京都外国語大学・外国語学部・講師
研究者番号：00434647