GDAに基づく統語情報付与XML化多言語平行資源の構築

堀 一成, 竹原 新†, 上原 順一‡, 小島 一秀∗, 藤家 洋昭†, 萬宮 健策†
大阪大学 大学教育実践センター, 世界言語研究センター†, 言語文化研究科‡,
サイバーメディアセンター∗

hori@cep.osaka-u.ac.jp, {takehara, k-mamiya}@world-lang.osaka-u.ac.jp,
uehara@lang.osaka-u.ac.jp, kkojima@cmc.osaka-u.ac.jp, huziieh@gmail.com

1　はじめに

我々、大阪大学の研究グループは、多言語の単語や会話文を一対一で比較できる表形式に整理した多言語平行資源の構築を進めてきた。これまで発表してきた会話文資源は、平文テキストの集積であったが、近年、GDA [1] に基づくXMLタグで表現した文の統語構造を含む資源の構築 [9, 10] を進めている。今回の発表では、資源構築の進行度合いには差があるが、中東・アジアの言語を含む資源の構築結果について報告する。

2　XML化されていない言語資源の構築

多数の言語を横断的に検索・比較できるようデータ化したものは、対照言語学・言語類型論を研究する際の基礎データとして意義のあるものだと考える。ここでは、単語集と、旅行会話文を中心とした会話文集を紹介する。多言語間の単語や文の対応で、各言語の対象とする語彙範疇は単純に一対一対応するものでないことは明らかである。しかし単純に一対一対応の表形式まとめることで、多言語間の対応関係調査を容易にする第一資料となるものを提供できるのではないかと考えている。

2.1　多言語単語集

単語集は、約5000語の単語を7言語並列し整理したものである。単語の選定基準は、日本語使用頻度順情報 [2] を参照し、その頻度上位のもののうち、日本語に固有で他言語で対応語を考えることが困難となるような語を除き、約5000語を対象とすることにした。対象言語は、アラビア語・ヒンディー語・ペルシア語・英語・中国語・朝鮮語・日本語である。その一部を図1に示す。

C	D	E	F	G	H	I
日本語	英語	ヒンディー語	ペルシア語	アラビア語	中国語	朝鮮語
ありがとう	thank you	धन्यवाद	متشكرم	شكرًا	谢谢	고마워
大きい	big, large	बड़ा	بزرگ	كبير	大	크다
おはよう	good morning	नमस्ते	صبح بخير	صباح الخير	您早	안녕
楽しい	pleasant	प्रसन्न	خوشحال	ممتع	愉快	즐겁다
あいさつ	greeting	नमस्कार	سلام	تحية	问候	인사
テレビ	television	टेलिविज़न	تلويزيون	تلفزيون	电视	텔레비젼
コーヒー	coffee	कॉफी	قهوه	قهوة	咖啡	커피
カメラ	camera	कैमरा	دوربين	آلة تصوير	照相机	카메라
父	father	पिता	پدر	أب	父亲	아버지

図1: 多言語5000単語集の一部

2.2　多言語会話文集

会話集は、旅行会話を中心とした約1000文を12言語並列し整理したものである。この文集は、大阪大学の教員を中心とする我々の研究グループが、各種会話集から選定し、多言語の翻訳が可能となるよう、各国独自の項目を改編したものである。文単位にID番号を付与し、エクセル表形式で保存している。対象言語は、アラビア語・スペイン語・英語・トルコ語・ヒンディー語・ペルシア語・日本語・モンゴル語・朝鮮語・中国語・ベトナム語・タイ語である。このうちの、スペイン語・英語・トルコ語・ペルシア語・日本語の５言語においては、会話文をネイティブ話者が吹き込んだ音声データも構築済みである [4]。その一部を図2に示す。

ID	日本語	英語	ペルシャ語	アラビア語	トルコ語	ヒンディー語
d_a_basic_001a	こんにちは。	Hello.	سلام.	السلام عليكم.	Merhaba.	नमस्ते/नमस्कार ।
d_a_basic_002a	おはようございます。	Good morning.	صبح بخیر.	صباح الخير.	Günaydın.	नमस्ते/नमस्कार ।
d_a_basic_003a	こんばんは。	Good evening.	شب بخیر.	مساء الخير.	İyi akşamlar.	नमस्ते/नमस्कार ।
d_a_basic_004a	おやすみ。	Good night.	شب بخیر.	ليلة سعيدة.	İyi geceler.	शुभ रात्रि ।
d_a_basic_005	はじめまして。	Nice to meet you.	از دیدن شما خوشوقتم.	فرصة سعيدة.	Tanıştığımıza memnun oldum.	आप से मिलकर बहुत ख़ुशी हुई ।
d_a_basic_006	自己紹介させてください。	Let me introduce myself.	اجازه بدهید خودم را معرفی بکنم.	دعني أقدّم لك نفسي.	Size kendimi tanıtayım.	अपना परिचय देता हूँ ।
d_a_basic_007a	さようなら。	Good bye.	خدا حافظ.	مع السلامة.	Hoşça kalın.	टा टा/फिर मिलेंगे ।
d_a_basic_008	気をつけてください。	Take care.	مواظب خودتان باشید.	اعتن بنفسك.	Kendine dikkat et.	अपना ख़्याल रखिए ।
d_a_basic_009a	すみません。	I'm sorry.	متأسفم.	أنا آسف.	Özür dilerim.	माफ़ कीजिए ।
d_a_basic_010	失礼ですが。	Excuse me.	ببخشید.	لو سمحت.	Affedersiniz.	ज़रा सुनिए ।

図 2: 多言語 1000 会話文集の一部

3 GDA に基づく XML 化言語資源の構築

前節で紹介した会話文のデータは、１文のテキストデータを単純に保存した、いわゆる平文データである。これに対して、言語教育の基盤データとなったり、自然言語処理システムの有用な参考情報となるために、統語情報を中心とする言語学的情報を付与した言語資源の作成がより重要であろうと考えている。本稿では、産業技術総合研究所で開発された GDA（大域文書修飾）に基づく情報を付与した会話文集資源について報告する。

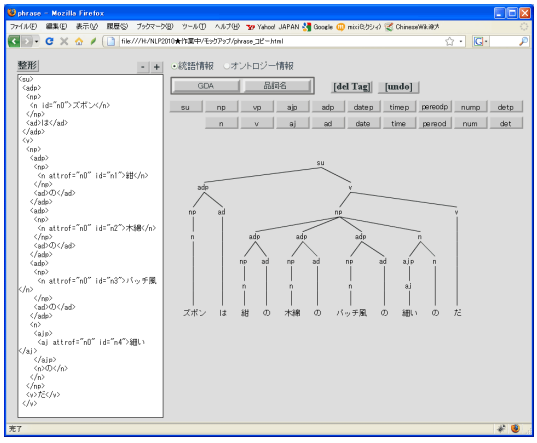


図 3: XML データ作成支援アプリケーション実行画面

3.1 XML データ作成作業をサポートするソフトウェア

紹介する XML データの作成は、外国語学を専攻する学生アルバイトに主に担当してもらっている。その作業は XMLSpy など、XML エディターで行っているのであるが、XML データに不慣れな者が多く、作業者の確保が困難である。少しでも作業者の心理的負担を軽減し、作業担当可能な者を増やす目的で、XML データ作成作業サポート Web ソフトウェアを開発してきている。これまでの言語処理学会等でその詳細は発表ずみである [5, 6, 8, 7]。開発したソフトウェアは、FLASH CS3 で開発 (Action Script のバージョンは 2.0) したものである。GUI 操作によって、木構造を画面上に作っていくことにより、GDA に準拠する統語情報を XML タグとして付与できる機能を持つ。このアプリケーションを Web ブラウザ上で実行し、文の木構造表示を行っている画面を図 3 に示す。

3.2 XML 化言語資源の例

現時点で、統語情報を付与したデータが完成、あるいはほとんど完成の状態にあるのは、日本語、英語、ペルシア語データである。トルコ語データは基本的な会話１００文程度の分量が完成している。

ここでは、XML データと木構造で示した結果の一例を図 4 に示すことで、成果の概要の提示とする。木構造の図は、前節で紹介したアプリケーションに、当該の XML データを入力することで、自動的に得られる画像をあてはめたものである。

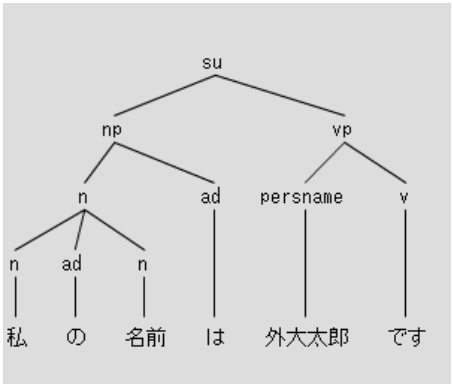
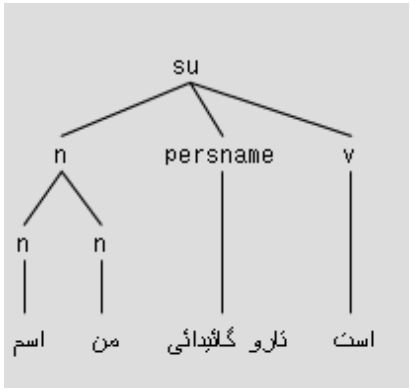
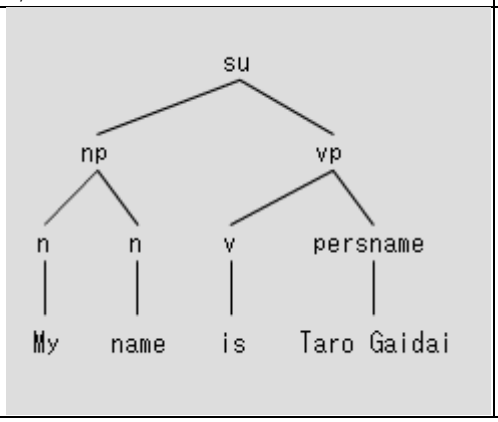
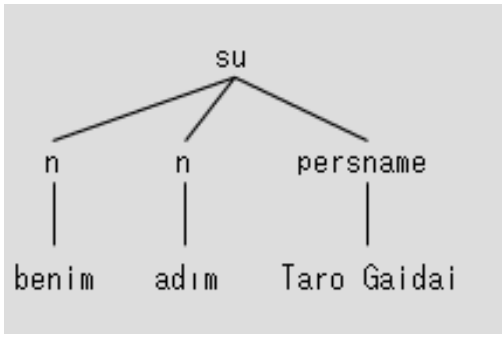
<table>
<tr><th></th><th>日本語</th><th>ペルシア語</th></tr>
<tr><td>GDA 構造化
XML データ</td><td><su syn="f">
<np><n syn="f">
<n>私</n>
<ad opr="arg">の</ad>
<n>名前</n></n>
<ad opr="aen">は</ad></np>
<vp>
<persname>外大太郎</persname>
<v>です。</v></vp>
</su></td><td><su syn="f">
<n syn="b" opr="aen">
<n>اسم </n>
<n>من </n>
</n>
<persname>تارو گائیدائی </persname>
<v>است </v>
.</su></td></tr>
<tr><td>データの
木構造表示</td><td> 
</td><td> 
</td></tr>
<tr><td></td><td>英語</td><td>トルコ語</td></tr>
<tr><td>GDA 構造化
XML データ</td><td><su>
<np opr="aen" syn="f">
<n>My</n>
<n>name</n>
</np>
<vp><v>is</v>
<persname>Taro Gaidai.</persname>
</vp>
</su></td><td><su>
<n>benim</n>
<n>adım</n>
<persname>Taro Gaidai.</persname>
</su></td></tr>
<tr><td>データの
木構造表示</td><td> 
</td><td> 
</td></tr>
</table>

図 4: GDA に基づく統語情報 XML データ化した多言語会話文集の一例

4　多言語資源の応用

多言語資源の活用のため、携帯端末で表示できるアプリケーションを開発してきている [3]。これは、開発した言語データのうち、自然災害等にみまわれた海外の被災地で救援活動を行う者に役立つ会話データがあり、これはそのデータを簡単に使えるよう工夫したものである。条件の悪い被災地で活用されることを想定し、語彙の選択方法や、発音カタカナ表記（現地語の専門家で無い日本人の利用を想定）、ゼスチャー映像の表示などが工夫点である。通常の Windows Vista, Windows7 上で動作するアプリケーションであるが、厳しい被災地環境での利用（Panasonic タフブックなどの携帯 PC）を想定したシンプルな操作となるよう工夫も行っている。

図 5: 携帯端末で災害救援用言語資源を表示するアプリケーション

5　おわりに

本稿では、大阪大学において、数年にわたり構築を続けてきている多言語資源の成果紹介を行った。本資源の構築開始時以降、さまざまな機関が多言語の言語資源の構築を発表している。我々の研究グループでは、それらの成果を参考に、各機関と協働して自然言語処理研究の発展の礎となり、また効果的な言語教育の素材となりうるような言語資源を構築していきたいと考えている。特に大阪大学　外国語学部の専攻言語は日本語を含め 25 言語あり、その知的財産の集積となるよう言語数の増加を図っていきたいと考えている。

謝辞　本研究は、科学研究費補助金 基盤研究 (B) 課題番号:19300047 『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』(研究代表者:堀　一成) と、科学研究費補助金 基盤研究 (B) 課題番号:22320103 『多言語会話文・語彙データベース構築と異文化交流におけるその活用に関する研究』(研究代表者:萬宮　健策) の補助を受け推進したものである。

参考文献

[1] 大域文書修飾 global document annotation(GDA). http://www.i-content.org/gda/.

[2] 天野 成昭, 近藤 公久（編）. 『日本語の語彙特性』第 2 期. 三省堂, 2003.

[3] 平松 初珠, 石島 悌, 萬宮 健策, 山根 聡, 堀 一成 . 多言語会話文、語彙データを利用した災害救援者教育用アプリケーションの開発. 情報処理学会 第 72 回全国大会講演予稿集　第 4 分冊, pp. 469 – 470, 2010.

[4] 堀 一成, 山崎 直樹, 竹原 新, 小島 一秀. 多言語平行マルチメディア言語資源の構築. 言語処理学会 第 13 回年次大会発表論文集, pp. 768 – 771, 2007.

[5] 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 回年次大会発表論文集, pp. 265 – 268, 2008.

[6] 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 回年次大会発表論文集, pp. 172 – 175, 2009.

[7] 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論文集, 第 4 分冊, pp. 119–122, 2010.

[8] 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 回年次大会発表論文集, pp. 478 – 481, 2010.

[9] 山崎直樹. XML による文法研究論文の構造化. 漢字文献情報処理研究, 第 3 号, pp. 38–45, 2002.

[10] 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 : 関西大学, 第 17 号, pp. 111 – 125, 2009.