多言語平行マルチメディア言語資源の構築

堀 一成　山崎 直樹　竹原 新　小島 一秀

大阪外国語大学　外国語学部

{hori, ymzknk, takehara, kkojima}@osaka-gaidai.ac.jp

1.はじめに

自然言語処理研究の発展のために、言語資源の充実が重要であることは明白である。これまで、多数の言語資源が構築され、様々に応用されることによって研究進展がはかられている。しかし、わが国において整備されてきた言語資源の対象となる言語は、日本語・英語・中国語・朝鮮語が主であり、加えてタイ・インドなど他のアジア諸語の資源構築が進められはじめたという状況である。世界中の主要言語をカバーするには程遠い段階であるといえる。

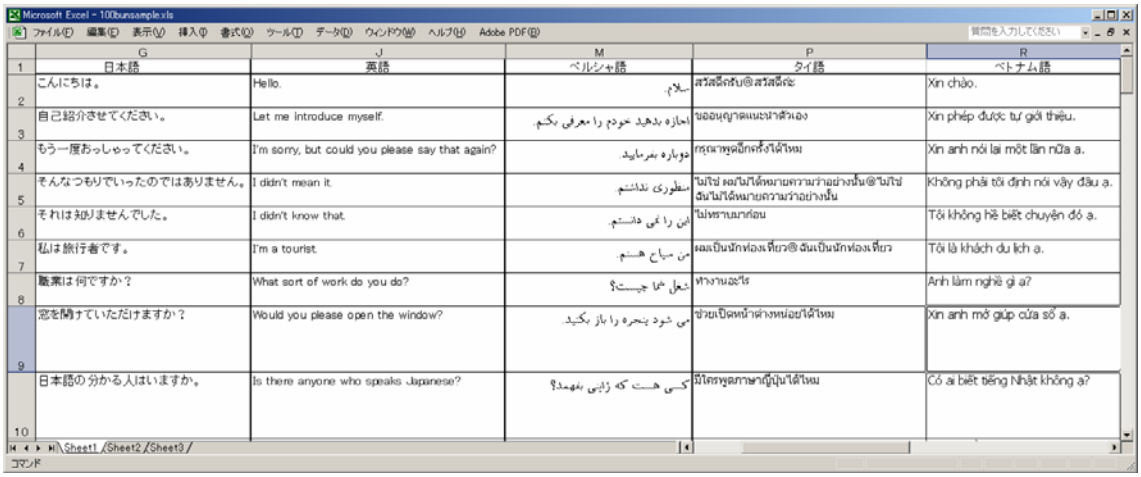
このような現況に対して、日本語を含む 25 言語を専攻語としている大阪外国語大学が、所属する研究者が持つ知見を結集して、多数の言語を横断的に検索・比較できるようデータベース化したものを構築すべく作業を進めている[1-3]。現時点において、13 言語の会話文をサンプルとして提供できることになり、本稿において報告するものである。

本稿で報告する言語資源の特徴は、以下のように考えられる。

- 既存の言語資源の活用で問題になるのは、権利関係である。本データはすべて大阪外国語大学のオリジナルデータであるので、利用に当たって問題が発生しない。また広く利用されるよう、構築資源の一部をフリーで公開する。
- さらに言語研究への応用（とくに言語類型論）と言語教育への活用（母語と目標言語との対照分析）を主な利用目的としている。
- また、一般に単語や文のカバーする範囲は言語により異なり、それを一対一で対応づけることは不可能なのであるが、この資源ではあえて一対一の対応による言語横断参照性を高める試みを行った。

2.多言語平行マルチメディア言語資源

大阪外国語大学においては、2000 年以来、前記の目的のもとに言語資源の構築を試みている。その方針は、大阪外国語大学の専攻語である 25 言語での構築を目標とし、しかし手作業による翻訳に基づくため作業能力の限界から各言語ごとのデータ量を絞っ



	G 日本語	J 英語	M ペルシャ語	P タイ語	R ベトナム語
2	こんにちは。	Hello.	سلام.	สวัสดีครับ@สวัสดีค่ะ	Xin chào.
3	自己紹介させてください。	Let me introduce myself.	اجازه بدهید خودم را معرفی بکنم.	ขออนุญาตแนะนำตัวเอง	Xin phép được tự giới thiệu.
4	もう一度おっしゃってください。	I'm sorry, but could you please say that again?	دوباره بفرمایید.	กรุณาพูดอีกครั้งได้ไหม	Xin anh nói lại một lần nữa a.
5	そんなつもりでいったのではありません。	I didn't mean it.	منظوری نداشتم.	ไม่ใช่ ผมไม่ได้หมายความว่าอย่างนั้น@ไม่ใช่ ฉันไม่ได้หมายความว่าอย่างนั้น	Không phải tôi định nói vậy đâu a.
6	それは知りませんでした。	I didn't know that.	این را نمی دانستم.	ไม่ทราบมาก่อน	Tôi không hề biết chuyện đó a.
7	私は旅行者です。	I'm a tourist.	من سیاح هستم.	ผมเป็นนักท่องเที่ยว@ฉันเป็นนักท่องเที่ยว	Tôi là khách du lịch a.
8	職業は何ですか？	What sort of work do you do?	شغل شما چیست؟	ทำงานอะไร	Anh làm nghề gì a?
9	窓を開けていただけますか？	Would you please open the window?	می شود پنجره را باز بکنید.	ช่วยเปิดหน้าต่างหน่อยได้ไหม	Xin anh mở giúp cửa sổ a.
10	日本語の分かる人はいますか。	Is there anyone who speaks Japanese?	کسی هست که ژاپنی بفهمد؟	มีใครพูดภาษาญี่ปุ่นได้ไหม	Có ai biết tiếng Nhật không a?

図１　多言語平行マルチメディア資源の一部（会話文テキスト情報）

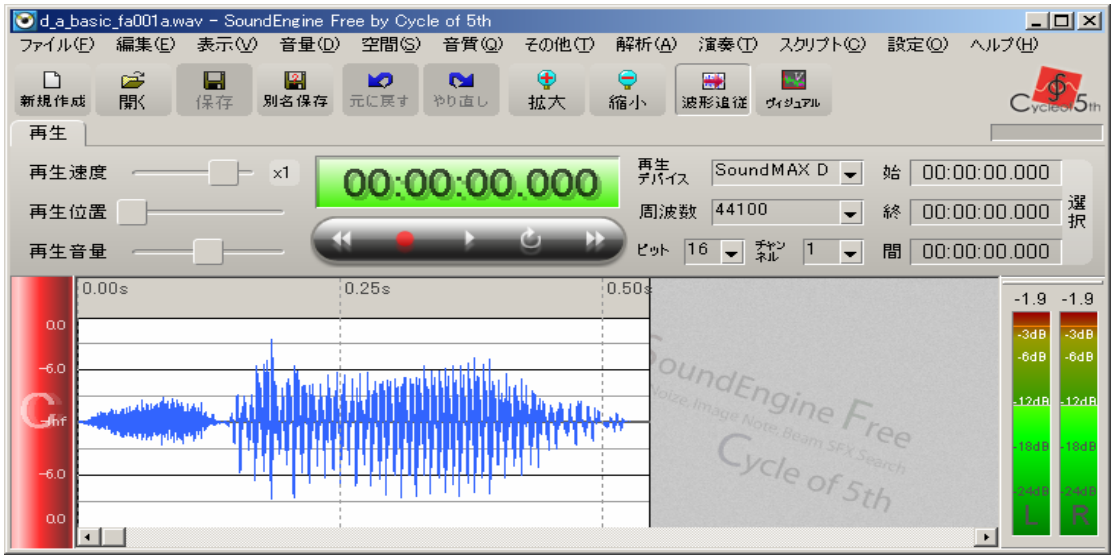


図２　多言語平行マルチメディア資源の一部（発話音声データ）の波形表示

たものとした。

その結果、主に旅行会話を想定した会話文約 1000 文（13 言語）と日本語使用頻度の高い単語［6］約 5000 語についてその翻訳を集めた単語集（７言語完成、4 言語作業中）、が作成できた。日本語と英語をキーとし、各言語に翻訳することで作業を進めた。また翻訳済みの単語・会話文の（主に大阪外国語大学のネイティブ教員による）発話音声データを録音し、文字データと対応付けてデータ化した。

この会話集のうち、利用頻度が高いと考えられる 100 文の文字データとその発話音声データを選定し、フリー言語資源として社会に提供すべく作業を進めている。

表１　構造化例文集のために選定した例文情報の一部（言語学的特長が付与されている）

二重主語	#026：	ゾウは鼻が長い	題目と主語をもつ文／二重主語文／大主語・小主語をもつ文	題目と主語が[全体−部分]の関係
二重目的語	#029：	わたしは彼に辞書を一冊送った	授与動詞をもつ文	二重目的語文になる？
受身	#045：	うちのロバがトラックにはねられた	受動（直接）	典型的な「被害」の意味をもつ受身
出現	#060：	この屋敷はよくお化けが出る	超自然現象の出現を表す文	一般の文と語順が異なるかどうか
結果	#066：	彼女は夫の父を思わずフライパンで殴って死なせてしまった	動作とその結果を表す構文	非意図的な結果のばあい

3.構造化例文集

前述の言語資源の構築と平行して、特に言語教育上重要と思われる、あるいは対照言語学的にみて重要と思われる文を 100 文選定し、その言語構造を GDA [5] (ただし一部独自の修正・拡張を行っている)によりタグ付けする試みも進めている。現時点では日本と中国語の作業を進めている。

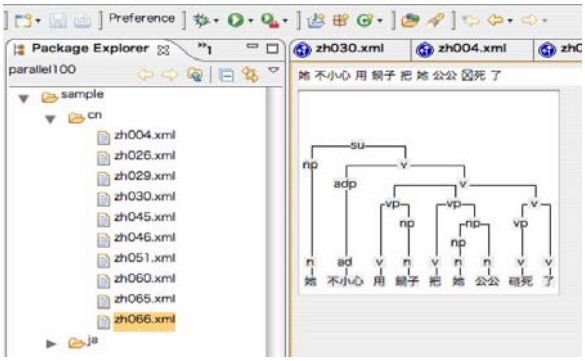


図３　タグ付け作業ソフトウェアの画面

これは、蓄積した言語資源を単に語学教育の基盤となる例文集として扱うだけでなく、タグ付けにより構造化された例文から文法規則を帰納により推測させたり、隣接言語との比較により当該言語の特徴を発見させるといった、新しいタイプの言語教育が可能になると考えている。また機械翻訳のシステム開発においても有用な参考情報となると考えている。表１に、その選定文（日本語）の一部を挙げる。選定に際してその文のどのような言語学的特徴に着目したかも付記している。

タグ付け作業にあたっては、東京工業大学が開発を進めている eBonsai [7] のシステムを利用し作業をおこなっている。その作業時の画面キャプチャを図３に示す。このような作業の出力として図４に示す XML

化テキストデータが得られる。

```
<?xml version="1.0" ?>
<data>
  <su>
    <np>
      <n>她</n>
    </np>
    <v>
      <adp>
        <ad>不小心</ad>
      </adp>
      <v>
        <vp>
          <v>用</v>
          <np>
            <n>鍋子</n>
          </np>
        </vp>
        <vp>
          <v>把</v>
          <np>
            <np>
              <n>她</n>
            </np>
            <n>公公</n>
          </np>
        </vp>
        <v>
          <vp>
            <v>砸死</v>
          </vp>
          <v>了</v>
        </v>
      </v>
    </v>
  </su>
```

図４　得られた構造化例文データ

4．おわりに

本稿では大阪外国語大学が構築を進めている言語資源について報告した。資源の一部は、言語資源協会を通じてフリーな利用に供する予定である。利用者の意見をフィードバックし、より有用な言語資源となるよう、資源の種類や構造化方法の検討をすすめる。今後は、テキストデータの構造化と言語数の拡充を推進する。

謝辞　本研究は 2006 年度大阪外国語大学特別研究費 II の補助を受けて進められたものである。

参考文献

[1] 堀一成 石島悌,「PostgreSQL による多言語単語データベースの構築」情報処理学会第 62 回全国大会講演論文集(2), (2001), pp.297-298.

[2] 堀一成 前田彩 石島悌「PostgreSQL と JSP を用いた多言語データベース検索アプリケーションの構築」FIT2002 一般講演論文集(2), (2002), pp.95-96.

[3] 堀一成「大阪外国語大学の言語資源を用いた言語 e-learning の構想」言語処理学会第 10 回年次大会ワークショップ「e-Learning における自然言語処理」論文集, (2003), pp.13-16.

[4] 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, (2002), pp.38-45.

[5] 「大域文書修飾 Global Document Annotation(GDA) 」 http://www.i-content.org/gda/

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

[7] 野口正樹 市川宙 橋本泰一 徳永健伸「構文木付きコーパス作成支援統合環境 eBonsai の新しいインターフェース」言語処理学会　第 12 回年次大会発表論文集, (2006) pp.751-754.