様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年 6 月 18 日現在

機関番号：３４３１５
研究種目：基盤研究（Ｃ）
研究期間：2009～2011
課題番号：21500151
研究課題名（和文）　日越パラレルコーパスとその日本語教育への応用

研究課題名（英文）　Japanese-Vietnamese Parallel Corpus and Applications to eLearning for Japanese Language

研究代表者　池田　秀人（IKEDA HIDETO）
立命館大学・情報理工学部・教授
研究者番号：30033905

研究成果の概要（和文）：
本研究は、急増するベトナムからの留学生に、日本語教えるための、教材を作成するための支援をすることが目的であった。その目的を達成するために、日英並列コーパスを作成し、日本の文献で頻出する用語や表現を探し出せる表現辞典を作成することにした。
研究成果としては、日英対訳コーパスとして、6028文対を入力して、日本語文を、依存性を使ったフレーズ関数という独自の方法で、その構造を解析すると同時に意味解析もして、注釈付コーパス(annotation corpus)として完成させた。また、これに対応するベトナム語文の構造も注釈をつけた。
また、ここで登録したフレーズだけを使って、ベトナム語文を作成すると、日本語文が出力される、「ベトナム語による日本語文書作成支援システム」を完成させた。また、この辞書を使ったベトナム語構文解析システムも実装した。

研究成果の概要（英文）：
The objective of this research is to prepare various educational materials of Japanese language for ballooning Vietnamese students　In order to attain the objective, the project developed Japanese-Vietnamese parallel corpus and the expression dictionary for users to find words and expressions of Japanese that frequently appeared in documents in Japanese.
As an result of the research, we inputted Japanese-Vietnamese corpus of 6028 sentences and developed a para-phrase dictionary with semantic annotations.
As the 1-st application of the para-phrase dictionary of Japanese and Vietnamese, the project developed an input support system of Vietnamese sentences that can be translated Japanese sentences with high quality.　As the 2nd application of this para-phrase corpus, I developed a parser of Vietnamese sentences.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,600,000	480,000	2,080,000
2010 年度	900,000	270,000	1,170,000
2011 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：自然言語処理

科研費の分科・細目：情報学、知能情報学
キーワード：日英パラレルコーパス、日本語教育、 機械翻訳、日本語作文支援、ベトナム語対訳フレーズ文法、多言語ｅラーニング

１．研究開始当初の背景

近年、IT 業界では安価な労働力を求めて、オフショア開発をすることが多くなってきている。1990 年代はインドがその対象国であったが、これが 2000 年代には中国に移り、現在ではベトナムがその対象国として注目を浴びている。

研究代表者はそれまで、アジアにおける IT 人材育成[1,2]のために、ベトナム人に情報技術や日本語教育を行うというプロジェクトに参加してきたが、ここでは、日本語教育と IT 関係の技術教育を、一定期間内（通常 4-6 年）に、必要な技能に到達させることが極めて困難な課題として浮かび上がっている。IT 技術を学ぶことだけでも大変な努力が必要なうえ、ベトナム人に日本語を習得させなければいけないという重い課題がある。そこで、IT 日本語という新しい科目が考え出されたが、まだ、学術的にも理論固めが不十分で、すぐに実施できる段階ではない。

研究代表者はこれまで、日本語コーパスの中から、IT 関係の文と新聞などの一般的日本語文との違いなどを調査し、IT 分野で使われる日本語の特徴を分析してきた。その結果、①語彙は専門用語を含むため大きく異なること、②文型は、システム仕様書や IT 分野の教科書、IT 専門雑誌などでは、「書き言葉」が多く、一般日本語教育で多くの時間を割いている「会話文型」は、比較的数が少ないこと、③IT 分野の文書をしっかり理解するためには、頻出する複文の構造と意味理解にもっと重点を置かなければいけないことなどがわかってきた。

また、ベトナムでの日本語教育の方法と実態を調査した結果、ベトナムでは、ほとんどの日本語学校や日本語コースが直接法（日本語で日本語を教える方法）で教えられており、これが日本語習得に時間のかかる大きな要因であることも分かってきた。ベトナム人の教師も日本語で教えている。これは、彼らも日本人の教師から日本語で教わったため、それをそのまま真似ているのである。

ベトナム語は現在のアルファベット表記を受け入れる前は、日本語と同じく漢字（「漢越語」という）で表記しており、多くの語彙が中国語に由来している。多くの語彙が日本語と共通の源をもっている。この特徴を積極的に使えば、ベトナム人が日本語を習得する時の１つの大きな障壁である「漢字用語」の問題が軽減される。

IT 用語は約 30%が純粋カタカナ用語で、「データベース管理」のような一部にカタカナを含んでいるものを加えると、70% が一部にカタカナを含んでいる。カタカナ用語の 98%は英語用語に対応しているので、英語の専門用語を知っていれば、IT 専門用語は大した障壁ではないことも、わかっている。

すると、残るは、機能語である。日本語教育は、学校文法をそのまま教えるのではなく、機能語という新しい文構成の単位を導入することで成功してきた。例えば、「の通りに」という用語は、学校文法では、「の（助詞）＋通（動詞）＋り（語尾）＋に（助詞）」の４つの品詞に分解できるが、これの文を構成する１つの単位（機能語）として教えるというものである。機能語は、他の言語でも考えられる。例えば、次の日本語文、ベトナム語文の組：

日本語文：*説明書の通りにやったがうまくいかなった。*

ベトナム語文：*Tôi đã làm theo như sách giải thích nhưng không tiến hành tốt.*

では、「の通りに」という日本語機能語が、「theo như」というベトナム語文の句に対応しているので、これは句ではなく、機能語と考えられる。しかし、実際はこのような単位を使って言語を教えているのは日本語だけである。そこで、ベトナム人にベトナム語で日本語文法の機能語をどのように教えるかという問題が浮上する。

この様な背景から、まだ確立していないベトナム語の機能語リストを作成し、日本語との対応をつける作業をしようと考えた。日本語の機能語は、日本語コーパスを分析することで作成でき　ることをすでに確認しているので、日本語・ベトナム語のパラレルコーパス（以後「JVPC」という。）を作成し、この作業をすることにした。

２．研究の目的

ここで確立したベトナム語の機能語を使って、「ベトナム語文意味分析法」、「ベトナム語の意味・表現関連」、「日本語・ベトナム語機能語対応」を確立し、すでに作成している「日本語文意味分析法」、「日本語の意味・表現関連」と合わせて、ベトナム人にとっての「日本文意味理解（読解法）」、および「日本語作文法」を確立することができる。

そしてこの２つの方法を使って、「ベトナム人に対する IT 日本語教育法」を確立しようというのが、この研究の目的である。「IT 専門用語教育法」、「日本文意味理解（読解法）」、「日本語作文法」を確立し、「ベトナム人のための IT 日本語テキスト」および、その関連教材を作成する。また、この方法による日本語教育を実際に試行し、その効果および今後の課題を明確にする計画である。また、日本語・ベトナム語のパラレルコーパスが確立すると、いろいろな応用が考えられる。そのため、データベースとして公開して、関連の研究に提供する。

３．研究の方法

研究の方法としては、「日本語表現辞典」のベトナム語訳を入力し、日本語―ベトナム語の並列コーパスを作成した。日本語はＯＣＲがあり、かなりいい進出で文字認識できたが、ベトナム語は人手による入力が避けられなかった。

次に、構文解析をする。日本語は、Cabochaという係り受け解析プログラムを使って、文のフレーズ分解および係り受け解析をしたものを使って、文をフレーズ分解した。

ベトナム語のフレーズ分解については、人手によっておこなった。これはベトナム語の構文解析プログラムの品質がまったく期待できないもので、係り受け解析プログラムもまだ開発されていない現状ではやむをえなかった。しかし、人手による解析を行ったため、日本語のフレーズに対応するよう、フレーズ分解されたため、日越フレーズの対応（アラインメント）が結果的になされるというメリットがあった。

こうしてできたフレーズ対を使って、並列フレーズ辞書を作成した。

また、辞書を使って、ベトナム語による日本語文書作成支援システムと、ベトナム語構文解析プログラムを作成した。

４．研究成果

日越対訳コーパスとして、6028文対を入力して、日本語文を、依存性を使ったフレーズ関数という独自の方法で、その構造を解析すると同時に意味解析もして、注釈付コーパス(annotation corpus)として完成させた。また、これに対応するベトナム語文の構造も注釈をつけた。

また、ここで登録したフレーズだけを使って、ベトナム語文を作成すると、日本語文が出力される、「ベトナム語による日本語文書作成支援システム」を完成させた。

更に、この成果をベトナム人に対する日本語教育に応用する実験も行った。実際、自分が言いたい（または、書きたい）文をベトナム語で入力すると、作成されたベトナム語文同時に、対応する日本語文も出力される。完成した辞書は、副詞句および副詞　733、名詞および名詞句　4140、述語構文　12103である。こでは、日本語能力検定試験の2級の出題範囲をほぼカバーしている。

この辞書を使って、機械翻訳を実装するため、対訳フレーズ辞書を使った構文プログラムを実装した。まだ、辞書の登録語・フレーズ集が少ないため、汎用翻訳システムをしては使えないが、日本語初級を学ぼうとするベトナム人学生にとって、有益なシステムであることが確認できた。

ここで、作成した対訳フレーズ辞書の構造のサンプルをしめす。

＜対訳フレーズ辞書サンプル＞

<対訳フレーズ, phraseNumber=000001>
<日本語フレーズ>
<日本語文>ステレオと本棚の間にテレビを置いた。</日本語文>
<構成フレーズ>
N0=私
N1=ステレオ
N2=本棚;
N3=テレビ;
P4=_と_の間に_を置く
([N1],[N2],[N3:object]);
S5=[_は]@v 連用タ接続:た
([N0:agent],[P4]){complete};
</<構成フレーズ>
</日本語フレーズ>
<ベトナム語フレーズ>
<ベトナム語文>Tôi đặt chiếc tivi ở giữa máy nghe nhạc và kệ sách. </ベトナム語文>
<構成フレーズ>
N0= Tôi;
N1= máy nghe nhạc
N2= kệ sách;
N3= tivi;
P4=chiếc-_-ở-giữa-_-và-_
([N3],[N1],[N2:object]);
S5= _-đặt-_([N0:agent],[P4]){complete};
</<構成フレーズ>
</ベトナム語フレーズ>
</対訳フレーズ>

この対訳フレーズの特長は、

- 文を関数型フレーズの列として表現していること
- 単語(N0~N3)、述語構文(P4)、文型(S5)もフレーズとしていること
- 省略（この場合 N0=私）もオプションとして記述されていること、
- 述語の活用形を（この場合、P5 内の ”@v 連用タ接続”）埋め込むことを示していること、
- フレーズ関数の引数に意味タグ（この場合、object, agent）を追加していること、
- 文型関数にテンス、アスペクト(この場合 S5 の complete)、およびモダリティを追加していること、

である。

また、図１に今回開発してベトナム語入力支援システムでホームページを作成した例を、図２に作成されたホームページの例を示す。実用に耐えられる日本文が生成されていることがわかる。

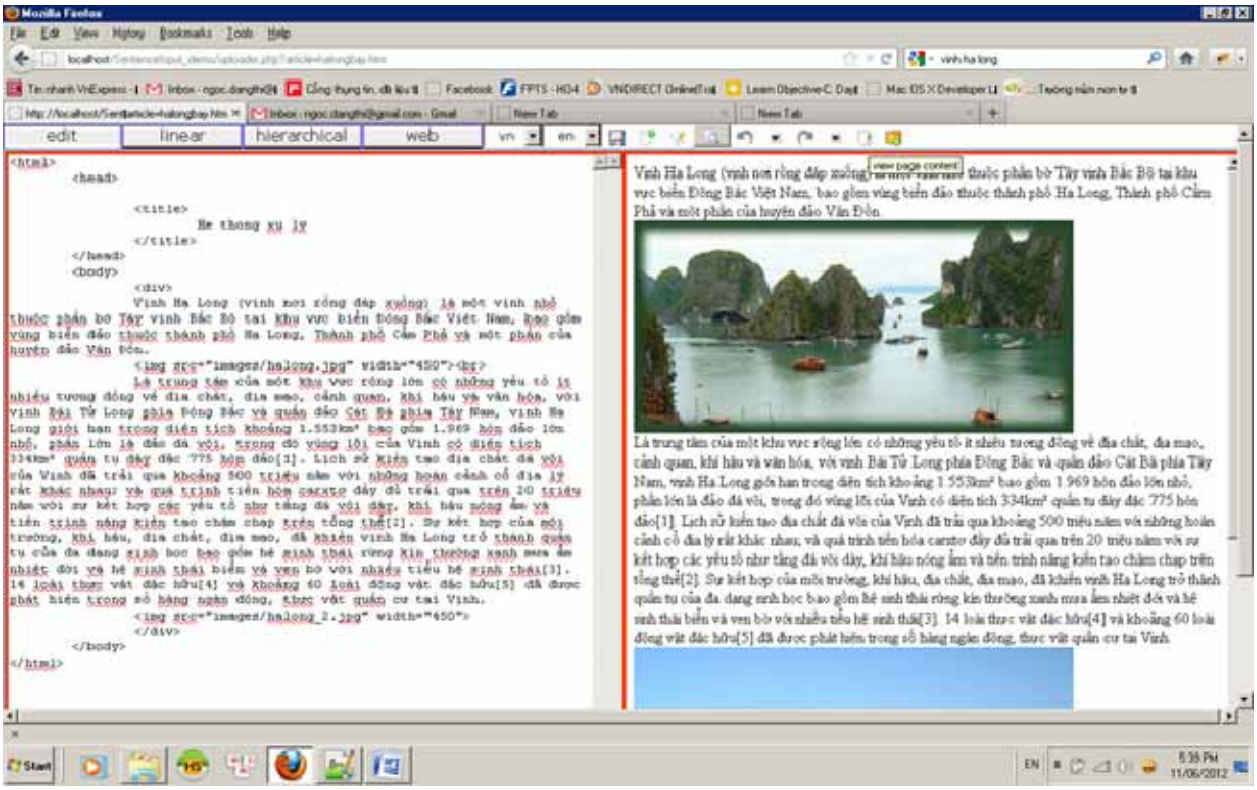

図１　ベトナム語入力支援システムで作成したホームページ

図 2-2．日本語翻訳されたホームページ

図 2-1．作成したベトナム語ホームページ

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）

〔学会発表〕（計　4 件）

① Ikeda,H. et.al　関数型フレーズ文法に基づくベトナム語構文解析プログラム開発，自然言語処理学会 2012 年 3 月 9 日，広島市立大学（広島県）

② Ikeda,H. et.al　Digitalization of Mathematical Text books and Evaluation of the Technolonogy, Proceedings of International Workshop on Digitalization of Mathematics (DEIMS12)　2012 年 2 月 16 日，発表場所（東京都）

③ Ikeda,H. et.al　Improved Word Alignment in Patent Domain, Proceedings of International Workshop on Patent Translation (MT-SUMMIT) 2011 年 10 月 5 日　厦門、中国

④ 池田秀人：数学教育のための多言語 e-Learning システム，「科学情報の電子化・自動処理・アクセシビリティをめぐる諸問題」　2011 年 2 月 12 日　筑波技術大学（茨城県）

〔産業財産権〕
○出願状況（計 1 件）

名称：自然言語文変換装置、自然言語文変換方法および自然言語文変換プログラム
発明者：池田秀人
権利者：（有）サイバープロ
種類：特許
番号：特願２０１０－０４１８１
出願年月日：2010 年 3 月 2 日
国内外の別：国内・米国

６．研究組織
(1)研究代表者
池田秀人（IKEDA HIDETO）
立命館大学・情報理工学部・教授
研究者番号：**30033905**