跡見学園女子大学文学部紀要　第 46 号　（2011 年 3 月 15 日）

日独パラレルコーパスを利用した用例集の作成

Creation of the example collection using the Japanese-German Parallel Corpus

阿部　一哉
ABE Kazuya

要旨

日本語の表現を外国語で表現したいという学習者のニーズに応えることは、外国語教育の根本的な課題の一つであろう。この課題を達成する一つの方法として、複数の言語で同じ内容のテキストを収集した言語資料であるパラレルコーパスを使用した用例集を作成することが有効であると考えられる。この想定に基づき日独対訳用例集の作成を行ったが、その際、形態素解析やデータ組み替え作業など、自然言語処理の分野で培われた手法を援用した。本研究は、言語研究で得られた成果を言語教育という形で社会に還元するために、情報技術を援用する、という研究モデルの一事例として位置付けられる。

0．はじめに

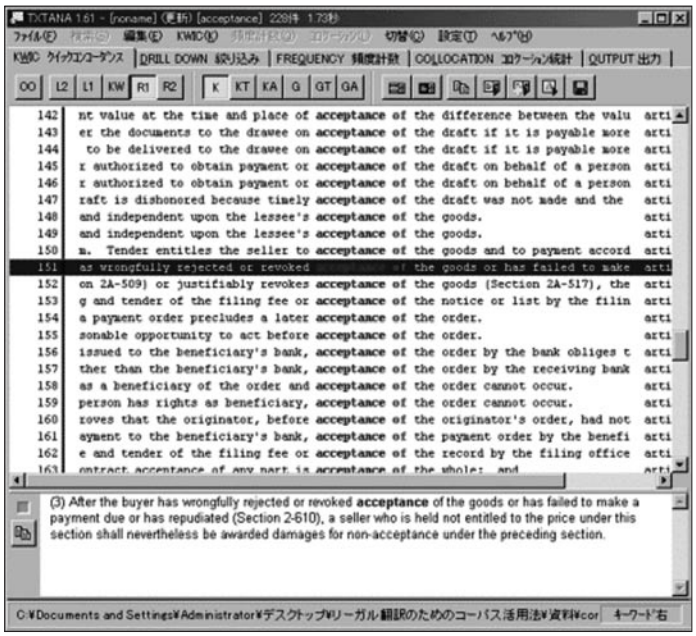
電式計算機のダウンサイジングとインターネットの普及によって、外国語教育の分野においても新たな教授法が提案されている。データ駆動型学習法（Data Driven Learning）では、（1）にあげるように、学習者は大量の言語資料をコンコーダンサと呼ばれるプログラムを使って閲覧し、帰納的に規則的パターンを抽出し、文法規則や語法といった要素を身につけていく。（Jones 1991）

次頁（1）はコンコーダンサで、acceptance が用いられている事例を KWIC と呼ばれる画面で提示し、右側の単語で並べ替えを行ったものである。学習者はこのデータ提示に基づき、acceptance は前置詞 of を伴って頻繁に用いられるという規則性を抽出することができるだろう。

和文独訳の指導においても、DDL のような発想は有効であると考えられる。

日本語の表現を外国語でどのように表現するのかを知りたい、というニーズに答えることは、どのように時代が変わろうとも外国語教授法がこたえなければいけないもっとも根本的な課題の一つであろう。本稿は、この課題を達成する一つの方法として、現在利用可能になったパラレル

(1)　コンコーダンサによる英文の閲覧画面



コーパス活用法 (2)（http://bgrass.halfmoon.jp/english/corpus/lc2/index.htm）より

コーパスと呼ばれる対訳データ集を利用することが有用であることを主張し、具体的に日独パラレルコーパスを用いて行った事例の整理について報告を行うものである。

まず第 1 章で、用例整理として具体的にどのようなものを想定しているかを述べ、第 2 章でパラレルコーパスの概要を述べる。第 3 章では、用例集作成における要件について見る。第 4 章では要件の具体的な解決方法を述べる。第 5 章では、作業と成果についていくつかの考察を加え今後の課題についてまとめる。

1．用例整理

冒頭でも述べたように、日本語の表現を外国語でどのように表現するのか知りたいという、学習者のニーズに応えることは、外国語教授法の最重要課題の一つと言える。学習者が、日本語の訳し方を習得するためには、日本語が外国語に訳されている事例に数多く触れる必要がある。これには、例えば、ドイツ語で書かれた小説とその邦訳を対照しながら見てゆく、といった学習法が、従来行われてきたものだろう。

対訳データを用いた学習において、分析的な手法を採るならば、(2) に挙げるように、ある一つの日本語の単語に着目し、その単語がドイツ語でどのように訳されているか見る、というものが考えられるだろう。

（2a）～（2c）では日本語表現と対応するドイツ語表現を対にした。「働く」と対応するドイツ語表現を下線で示した。

（2） 働く（はたらく）

a. しばしばふた親とも働かなければならない

Oft müssen beide Elternteile arbeiten.

b. 効率的な働き方が重要である

Eine effiziente Arbeitsweise ist wichtig.

c. 彼は大学教師としてベルリンで働いていた

Er war als Hochschullehrer in Berlin tätig.

（2a）では「働く」という日本語が、「～ねばならない」という表現の一部として用いられている。対応するドイツ語表現は müssen ... arbeiten である。

（2b）では「働く」という日本語が「～のし方」という表現の一部として用いられている。対応するドイツ語表現は Arbeitsweise である。

（2c）では「働く」という日本語に対して、tätig sein という他の 2 例とは異なる形式が使われている。

このように対訳データを用いることで、学習者は「働く」という日本語に arbeit- / tätig sein といった表現形式が対応していることだけではなく、これらのドイツ語表現が具体的にどのように用いられるのかも、知ることができるのである。つまり、以上のデータ提示方法により、学習者は日本語の訳し方を帰納的に学習することも可能であると考えられるのである。

ただし、こういったデータの提示も、事例件数が少なければ、あまり効果が期待できない。したがって、日本語の表現を外国語でどのように表現するのかを知りたいという、学習者のニーズに応えるためには、一つの日本語表現に対して、複数の外国語表現が対応している事例を提示することが有用であると考えられるのであるが、その際可能な限り数多くの用例を提示することが必須条件となる。

このために、本稿ではパラレルコーパスと呼ばれる対訳データからなる言語資料を利用する。第 2 章では、このパラレルコーパスについて見る。

2．パラレルコーパスとは何か

前章では、日本語の単語に対してドイツ語の翻訳事例を複数提示することで、「訳し方」を帰納的に発見する可能性を示唆した。本稿では、そのような教材を作成するためにパラレルコーパスの使用を提案する。本章ではこのパラレルコーパスについて見ていく。

2. 1 パラレルコーパスの概要

パラレルコーパスとは（3）に挙げるように、同じ内容をもつテキストを複数の言語で収集した言語資料である。

（3） a. Das ist ein Kugelschreiber. Der Kugelschreiber ist meins.

b. これはボールペンです。このボールペンは私のものです。

（3a）はドイツ語のテキスト、（3b）は日本語のテキストで、同じ内容を表している。

実際にパラレルコーパスを用いる場合、（4）のように、文を対応付けたり、（5）のように、単語を対応付けたりといった作業を、用途に合わせて行う。

（4） a. Das ist ein Kugelschreiber.　これはボールペンです。

b. Der Kugelschreiber ist meins. このボールペンは私のものです。

（5）

Das	
ist	
ein	これは
Kugelschreiber	ボールペン
	です。

（4）では、それぞれ（4a）（4b）で、日独で同じ意味をあらわす文が対応付けられている。

（5）では日独で同じ意味をあらわす単語、Kugelschreiber「ボールペン」で、文が対応付けられている。

次節では、対訳データを集めた言語資料であるパラレルコーパスの用途について見る。

2. 2 パラレルコーパスの用途

パラレルコーパスは、言語研究の基礎資料として用いられる場合（たとえば山崎（2009））や、機械翻訳の基礎資料として用いられる場合（例えば Google のプロジェクト）などがある。

これらの用途と並んで、外国語教育の現場でもパラレルコーパスは利用されている。

外国語の運用能力の習得には、外国語で書かれたテキストや、会話の音声資料といった、いわゆるオーセンティックなデータを利用することが有効である。

しかし、実際オーセンティックな言語資料には、学習者が未習の表現や語彙なども使用されているため、そのまま教材データとして用いるには難易度が高すぎて、利用しにくい場合が多い。

この点、パラレルコーパスはオーセンティックな言語資料を、母語の訳付きで学習者に提供することができるため、学習者のテキスト理解が格段に容易になる。実際に冒頭でふれたデータ駆動型発見学習の実践事例においても、パラレルコーパスの有用性が確認され、一定の成果を挙げている。（中條 2005, 2007）

以上パラレルコーパスの概略と用途について見た。パラレルコーパスは、日本語の単語に対するドイツ語の用例集を作成するといった目的には最適な素材であろう。しかしそのためには、データを加工し日本語の単語と用例を関係付ける作業を行う必要がある。そこで第3章では、パラレルコーパスの加工作業において、満たさなければいけない要件について見てゆく。

3．作業上の要件

本章ではパラレルコーパスを本稿での用途に合わせ加工する際に、作業上満たさなければいけない要件について見てゆく。

利用するデータは、東京外国語大学大学院の在間研究室で作成した、言語研究用の対訳事例集（非公開）で、(6) に挙げるように、ドイツ語の事例と日本語の訳文がペアになっている。そのままパラレルコーパスと呼んでいい。

(6) Er ist gebürtiger Bayer. 彼は生粋のバイエルン人だ
sich fachkundig beraten lassen 専門知識に基づく助言を求める
die Nachhaltigkeit der europäischen Konjunktur ヨーロッパの景気の持続可能性

3. 1 データの加工作業

本稿では (7) に挙げる形式の対訳データをすでに第1章で見たような、(8) の形に組み替える。

(7) a. しばしばふた親とも働かなければならない
Oft müssen beide Elternteile arbeiten.
b. 効率的な働き方が重要である
Eine effiziente Arbeitsweise ist wichtig.
c. 彼は大学教師としてベルリンで働いていた
Er war als Hochschullehrer in Berlin tätig.

(8) 働く（はたらく）
しばしばふた親とも働かなければならない
Oft müssen beide Elternteile arbeiten.
効率的な働き方が重要である
Eine effiziente Arbeitsweise ist wichtig.
彼は大学教師としてベルリンで働いていた
Er war als Hochschullehrer in Berlin tätig.

(7) の 3 事例では下線で示すように「働く」という単語の活用形が用いられている。これらを「働く」という単語でまとめると、(8) になる。

3. 2 要件

(8) にあげるようなデータ提示を、可能な限り多くの単語について行うためには、パラレルコーパスの日本語訳に含まれるすべての単語について組み替え作業を行う必要がある。

また、学習者が単語の訳し方を調べるに際して、ただやみくもに日本語の単語が羅列されているというデータの提示方法では、使い勝手が悪い。そこで、今回は辞書と同じように日本語の単語を読みの 50 音順に配列して、以下 (9) の流れで提示する。一行目が学習者の挙動タイプ、二行目が「働く」を選択するまでの具体的な選択項目を示している。

(9)　語頭の文字が含まれる 50 音の行選択　→　語頭の文字選択　→　単語の選択
　　はひふへほ　　は　　働く（働く）

今回はデータの提示方法として、HTML ファイルからなる静的サイトを使用する。そのため (9) のような手順を選択した。その他さまざまな提示方法が考えられるだろうが、この議論については本稿とは別の場で行いたい。

まとめると、今回の加工作業では、(10) にあげる 5 つの要件を満たす必要がある。

(10)（要件 1.）パラレルコーパスの日本語に含まれるすべての単語を取り出す。
（要件 2.）取り出された単語一つ一つについての辞書形を得る。
（要件 3.）同じ単語が用いられている用例をまとめる。
（要件 4.）単語ごとにまとめられた用例を適切な方法で並べ替える。
（要件 5.）日本語の単語と対応するドイツ語表現をマークアップする。

このうち（要件 5.）「日本語の単語と対応するドイツ語表現をマークアップする。」というのは、具体的には次の（11）のことである。

(11) 働く（はたらく）
しばしばふた親とも働かなければならない
Oft müssen beide Elternteile arbeiten.

日本語の辞書形「働く（はたらく）」に対応しているのは、それぞれ下線でしめした、日本語の活用形「働か」、ドイツ語の arbeiten である。日本語の単語がどのような表現に訳されて、またその表現が具体的にどのように用いられているか知るためには、ドイツ語の事例で当該の単語が下線なり色を変えるなりの方法で示されているのが望ましい。しかし今回実際にはこの要件は満たされなかったのであるが、この問題については次の章で述べる。

以上、具体的な加工作業の内容と、その際満たさなければならない要件について述べた。第 4 章ではこれらの要件を解決する手法について見る。

４．解決方法

本章では、第３章でまとめた要件を満たすため、具体的にどのような解決方法をとるのかについて述べる。

今回使用したパラレルコーパスは1615文という量であった。これらの文の日本語訳について、すべての文を単語に切り分け、単語に文を振り分けて、単語ですべての事例をソートするのは、限られた時間の中では不可能である。今回はこの問題を解決するために、事例に対して形態素解析を施し、スクリプト言語のひとつであるpythonで処理を記述し、組み替え作業を行った。

以下4.1節で今回使用した形態素解析作業について、4.2節でデータの組み替え作業について、それぞれ述べる。

4. 1 形態素解析

形態素解析とは、テキストに含まれるすべての単語について、その品詞や語形変化のカテゴリといった形態情報を付与することである。この形態素解析を行うプログラムを形態素解析プログラムと呼ぶ。形態素解析を行うためには、テキストを文に区切り、文を単語に区切る機能が必要となる。今回はこの単語切りだし機能を主に利用する。

形態素解析プログラムとしてMeCabを利用した。MeCabは京都大学情報学研究科－日本電信電話株式会社コミュニケーション科学基礎研究所 共同研究ユニットプロジェクトを通じて開発されたオープンソース 形態素解析エンジンである。紙面の都合上詳述は避けるが他の日本語形態解析プログラムである、ChaSen, Juman, KAKASIよりも、機能や処理速度の点において優っている、とされている。（MeCabプロジェクトページ）

次にMeCabの今回の使用法について述べる。

まず、（12）に挙げるように、パラレルコーパス（12a）から（12b）のように日本語文のみをとりだす。

（12）a.　die Erarbeitung eines neuen Konzeptes
新しい構想の作成
Er ist im zivilen Leben Bauingenieur.
彼は市民生活では（除隊すれば）建築技師だ
Das ebnet uns neue Wege.
それは私たちに新たな道を開いてくれる
eine bessere Betreuung der Studierenden durch Tutoren
チューターによる学生のよりよい世話［をすること］、面倒をみること

b.　新しい構想の作成

彼は市民生活では（除隊すれば）建築技師だ

それは私たちに新たな道を開いてくれる

チューターによる学生のよりよい世話［をすること］、面倒をみること

次に、この日本語テキストに対して形態素解析を行う。

その結果、このテキストに含まれるすべての単語に関して（13）に挙げるような情報が付与される。紙面の都合上、「する」の活用形が含まれる二文目のみの結果について挙げる。

（13）彼　　名詞 , 代名詞 , 一般 ,*,*,*, 彼 , カレ , カレ
は　　助詞 , 係助詞 ,*,*,*,*, は , ハ , ワ
市民　名詞 , 一般 ,*,*,*,*, 市民 , シミン , シミン
生活　名詞 , サ変接続 ,*,*,*,*, 生活 , セイカツ , セイカツ
で　　助詞 , 格助詞 , 一般 ,*,*,*, で , デ , デ
は　　助詞 , 係助詞 ,*,*,*,*, は , ハ , ワ
（　　記号 , 括弧開 ,*,*,*,*, （, （, （
除隊　名詞 , サ変接続 ,*,*,*,*, 除隊 , ジョタイ , ジョタイ
すれ　動詞 , 自立 ,*,*, サ変・スル , 仮定形 , する , スレ , スレ
ば　　助詞 , 接続助詞 ,*,*,*,*, ば , バ , バ
）　　記号 , 括弧閉 ,*,*,*,*,）,）,）
建築　名詞 , サ変接続 ,*,*,*,*, 建築 , ケンチク , ケンチク
技師　名詞 , 一般 ,*,*,*,*, 技師 , ギシ , ギシ
だ　　助動詞 ,*,*,*, 特殊・ダ , 基本形 , だ , ダ , ダ

全体がタブ区切りで大きく二カラムに分けられている。左カラムのリストが解析前の文中形で、これらを解析すると右カラムの結果になる。右側のカラムでは解析して得られた諸特性が「,」で区切られ提示されている。前章に挙げた要件のうち、（要件 1.）「パラレルコーパスの日本語に含まれるすべての単語を取り出す。」が、この作業で満たされる。

次に辞書形の抽出について述べる。（13）の 9 行目の「すれ」は次の（14）のように解析されている。

（14）すれ　動詞 , 自立 ,*,*, サ変・スル , 仮定形 , する , スレ , スレ

右カラムが、「すれ」を解析した結果であり、諸特性が「,」で区切られていることについては既に見た。このうち右から 3 つ目が辞書形である。そこでこの特性のみを取り出して、用例をまとめるための見出し語として利用する。前章で挙げた要件のうち、（要件 2.）「取り出された単語一つ一つについての辞書形を得る。」というものがこの作業で満たされる。

すべての例文から抽出された辞書形はいったんリスト化し、重複項目を削除する。例えば、上

掲（13）のリストからは、以下（15）の辞書形リストが得られる。

（15）全単語　　重複単語を削除したリスト

彼	彼
は	は
市民	市民
生活	生活
で	で
は	（重複しているので削除）
(	(
除隊	除隊
する	する
ば	ば
)	)
建築	建築
技師	技師
だ	だ
15 語	14 種類

「全単語」カラム 6 行目の助詞の「は」は、2 行目で既出により重複しているので、「重複単語を削除したリスト」カラムでは削除されている。この結果、処理を行った文「彼は市民生活では（除隊すれば）建築技師だ」から、14 種類の単語が得られた。

この作業を 1615 文からなるパラレルコーパス全体に関して行った結果 2824 種類の日本語単語が得られた。

これらの見出し語に対してはさらに作業を行ってゆくので、この時点で見出し語と用例の対応関係を記録しておく。のちほどこの対応関係を参照しながら用例の組み替えを行ってゆく。

前章の要件のうち、（要件 4.）「単語ごとにまとめられた用例を適切な方法で並べ替える。」を満たすためには、単語に漢字が含まれるものについて「読み」を付与する必要がある。MeCab には単語の読みを付与する機能が実装されているが、この機能は独立して実行することも可能である。（16）はこの機能を用いて（16a）に含まれる単語に対して、読みの付与を行い、（16b）の出力が得られることを示したものである。なお、実際は改行区切りされた辞書形に対して、改行区切りされた辞書形の読みを出力させたが、紙面の都合上いずれも改行をスラッシュ（/）で示してある。

（16）a.　彼 / は / 市民 / 生活 / で / は /（/ 除隊 / する / ば /）/ 建築 / 技師 / だ

b.　カレ / ハ / シミン / セイカツ / デ / ハ /（/ ジョタイ / スル / バ /）/ ケンチク / ギシ / ダ

この機能を使い、辞書形リストの辞書形一つ一つに対して読みを付与した。（17）に挙げる「者→モノ」ような誤読も含まれるが、この点の修正については次回の課題としたい。

（17）a.　容疑 / 者 / は / …

b.　ヨウギ / モノ / ハ / …

以上の作業の結果、パラレルコーパスに含まれる日本語の「単語辞書形リスト」「単語辞書形読みリスト」が得られ、前章で挙げた要件のうち、（要件 1.）（要件 2.）（要件 4.）が満たされた。次節ではこれらのデータを用いてさらに行った作業について述べる。

4. 2 組み替え作業

4.1 節では、形態素解析プログラム MeCab を用いて行った作業について述べた。この作業で「単語辞書形リスト」「単語辞書形読みリスト」が得られた。次にこれらのデータを用いて、組み替え作業を行う。

MeCab の仕様から、単語の読みはカタカナで出力される。しかし見出し語の読みとしてはカタカナではなく、ひらがなを用いるのが普通である。これは日常的な言語使用の感覚とも一致しているだろう。そこでまず「単語辞書形読みリスト」を平仮名化する必要がある。

また、「単語の辞書形」「単語の辞書形読み」はデータを組み替える際、見出し語の役割を担う。日本語を見出し語とした辞書特有の問題として、並べ替えの際の濁音や引きの問題がある。今回は、文字コードは UTF-8 を使用したが、UTF-8 の仕様では、濁音・引きは通常の仮名の後に配置されているため、その順番に従って並べ替えを行うと、（18）のように、日常的な言語使用の感覚（18a）とは異なる（18b）のような並べ替えが行われてしまう。

（18）a.　日常的な言語使用の感覚に従った配列

ヒイジイサン　ビーズ　ビイドロ　ヒルマ　ビルマ

b.　実際の配列

ヒイジイサン　ヒルマ　ビイドロ　ビルマ　ビーズ

そこで、辞書形に対して次の（19）の操作を行った。

（19）辞書形（読みカタカナ）

↓　[平仮名への変換]

辞書形（読み平仮名）

↓　[ソートキーへの変換・ソートキーの追加]

辞書形（読み平仮名・読みソートキー）

変換では辞書形の読みについて必要に応じて一文字一文字を変換する作業を行った。その際変換テーブルを使用するが、具体的に、「平仮名への変換」では（20）、「ソートキーへの変換」では（21）に挙げるような変換表を使用した。

（20）ア → あ , イ → い , ウ → う , エ → え , オ → お ,
カ → か , キ → き , ク → く , ケ → け , コ → こ ,
サ → さ , シ → し , ス → す , セ → せ , ソ → そ ,
タ → た , チ → ち , ツ → つ , テ → て , ト → と ,
ナ → な , ニ → に , ヌ → ぬ , ネ → ね , ノ → の ,
ハ → は , ヒ → ひ , フ → ふ , ヘ → へ , ホ → ほ ,
マ → ま , ミ → み , ム → む , メ → め , モ → も ,
ヤ → や , ユ → ゆ , ヨ → よ , ラ → ら , リ → り ,
ル → る , レ → れ , ロ → ろ , ワ → わ , ヲ → を ,
ン → ん
ガ → が , ギ → ぎ , グ → ぐ , ゲ → げ , ゴ → ご ,
ザ → ざ , ジ → じ , ズ → ず , ゼ → ぜ , ゾ → ぞ ,
ダ → だ , ヂ → ぢ , ヅ → づ , デ → で , ド → ど ,
バ → ば , ビ → び , ブ → ぶ , ベ → べ , ボ → ぼ ,
パ → ぱ , ピ → ぴ , プ → ぷ , ペ → ぺ , ポ → ぽ ,
ァ → ぁ , ィ → ぃ , ゥ → う , ェ → ぇ , ォ → ぉ ,
ャ → ゃ , ュ → ゅ , ョ → ょ ,
ヴ → ヴ , ッ → っ , ヰ → ゐ , ヱ → ゑ , ー → ー , [→ [,] →] , 〈 → 〈 , 〉 → 〉 ,

（21）a. 引きの処理

あー→ああ	かー→かあ	さー→さあ	たー→たあ
なー→なあ	はー→はあ	まー→まあ	やー→やあ
らー→らあ	わー→わあ	がー→があ	ざー→ざあ
だー→だあ	ばー→ばあ	ぱー→ぱあ	ぁー→ぁあ
ァー→ァあ	ゃー→ゃあ	えー→ええ	けー→けえ
せー→せえ	てー→てえ	ねー→ねえ	へー→へえ
めー→めえ	れー→れえ	げー→げえ	ぜー→ぜえ
でー→でえ	べー→べえ	ぺー→ぺえ	ぇー→ぇえ
ェー→ェえ	おー→おお	こー→こお	そー→そお
とー→とお	のー→のお	ほー→ほお	もー→もお
よー→よお	ろー→ろお	をー→をお	ごー→ごお

ぞー→ぞお　どー→どお　ぼー→ぼお　ぽー→ぽお
ぁー→ぁお　ょー→ょお　うー→ううくー→くう
すー→すう　つー→つう　ぬー→ぬう　ふー→ふう
むー→むう　ゆー→ゆう　るー→るう　ぐー→ぐう
ずー→ずう　づー→づう　ぶー→ぶう　ぷー→ぷう
ぅー→ぅう　ゅー→ゅう　いー→いい　きー→きい
しー→しい　ちー→ちい　にー→にい　ひー→ひい
みー→みい　りー→りい　ぎー→ぎい　じー→じい
ぢー→ぢい　びー→びい　ぴー→ぴい　ぃー→ぃい

b.　濁音・半濁音の処理

が→か　ぎ→き　ぐ→く　げ→け　ご→こ
ざ→さ　じ→し　ず→す　ぜ→せ　ぞ→そ
だ→た　ぢ→ち　づ→つ　で→て　ど→と
ば→は　び→ひ　ぶ→ふ　べ→へ　ぼ→ほ
ぱ→は　ぴ→ひ　ぶ→ふ　ぺ→へ　ぼ→ほ

以上の作業の結果、見出し語の候補である辞書形に対して、「辞書形読み平仮名」「辞書形ソートキー」が追加された。

次に、データの並べ替え、出力を行う作業について述べる。辞書形に対して、その辞書形が使用されている用例をまとめるので、最終的な出力は以下（22）に挙げるものになる。

（22）働く（はたらく）

しばしばふた親とも働かなければならない
Oft müssen beide Elternteile arbeiten.
効率的な働き方が重要である
Eine effiziente Arbeitsweise ist wichtig.
彼は大学教師としてベルリンで働いていた
Er war als Hochschullehrer in Berlin tätig.

既に述べたが、今回は単純な HTML ページからなる Web サイトの形でデータを出力する。そのサイトマップは以下のようになる。

（23）
```
/siteroot                          （サイトトップ）
├── a                              （50 音行インデックス）
│   ├── a                          （50 音インデックス）
│   │   └── sentences    （単語ごとの文）
```

```
|        …                                    （中略）
|   └── u
|        └── sentences
…                                             （中略）
```

このために、見出し語の読みに従って、見出し語を 50 音別に振り分けて、50 音をローマ字に変換しディレクトリに配置する、という作業を行う必要がある。これらの処理には、上で既に見た変換テーブルを用いた処理が応用できる。概略データの組み替えと出力は以下（24）の手順で行った。

（24）「辞書形ソートキー」の語頭文字への振り分け

↓

「辞書形ソートキー」に従った「見出し語」の並べ替え

↓

「見出し語」と用例の対応付け

↓

データの出力

その結果出来上がった Web ページは以下（25）のとおりである。

（25）

サイトトップ

50 音行インデックス：あ行

ひらがなインデックス：あ

単語インデックス：上げる（あげる）

学習者はサイトルートで調べたい単語の語頭の文字の行を選択し、次に語頭の文字を平頭の文

字の行を選択し、次に語頭の文字を平仮名で選択し、最後に単語を選択し単語インデックスページを閲覧する。このようなインターフェースを用いて、日本語の単語からドイツ語の文を閲覧することを可能にした。

なお、理想的には次の（26）に挙げるように、日本語の単語と対応するドイツ語表現がマークアップされるべきであることには既にふれた。

（26）働く（はたらく）

しばしばふた親とも働かなければならない

Oft müssen beide Elternteile arbeiten.

効率的な働き方が重要である

しかしこれを行うためには、パラレルコーパスにおける「単語アラインメント」という問題を解決しなければいけない。単語アラインメントとは、意味的に対応する二言語のテキスト間で、意味的にも位置関係的にも一致している単語同士を対応づけることである。この「単語アラインメント」は自動化するためには、日本語とドイツ語の単語対応データなどを利用し、統計処理を行う必要があり今回は断念した。今後の課題としておきたい。

5．考察

今回 1615 文のデータを加工して、2824 種類の見出し語からなる用例集を作成することができた。日本語から引ける訳付きドイツ語用例集（以下便宜上「用例集」とする）の作成という目標は達成できたが、解決するべき問題点も見えてきた。以下 3 点述べる。

問題の 1 点目は用例数の問題である。単語によっては助詞の「は」のように 400 事例以上ヒットするものもあれば、「買収（ばいしゅう）」のように 1 事例しかヒットしないものもある。どのような用例数が適切なのかという点は、学習者評価を行いさらなる考察を行う必要がある。またリソースとなるパラレルコーパスの規模を大きくして、必要に応じて用例数を操作することができるようにしておく必要がある。

問題の 2 点目は「用例集」の使い勝手の問題である。他のデータ駆動型学習では本稿冒頭で見たように、コンコーダンサというプログラムを介してデータが提示される。このプログラムは別途新たにインストールする必要があり、この作業が学習者には不評という意見もある（中條 2005, 2007）。これに対して今回必要なプログラムはインターネット・エクスプローラなどの Web ブラウザで、インターネットに接続できる環境にある学習者であれば、新たなインストール作業を行う必要はない。今後も可能な限り Web ブラウザでデータ提供を行うという方針は変えないでいたほうがよいだろう。なお、コンコーダンサを Web ブラウザで実装するという試みも行われており（たとえば Corpus Concordance English）、こういった成果も随時参考にしていきたい。

なお、今回作成した「用例集」のデータの提示方法は、「50 音行選択→平仮名選択→単語選択」という 3 回のページ遷移を経なければ求めるデータにたどり着くことができない。また、Google の提供しているような検索機能を実装することも考えられる。「用例集」の使い勝手に関しては、今後学習者評価を参考にしながら改良してゆく必要がある。

問題の 3 点目は、データ操作の問題である。既に述べた問題として、日本語単語とドイツ語単語の対応づけを、今回は行うことができなかった。可能であれば対応付けは行ったほうがいいのだろうが、その一方で日本語単語に対する用例数が増えれば、単語の対応付けも帰納的に推論できるという可能性も考えられる。学習者評価に基づきつつ、適正な用例数の問題と絡めながら、考えていきたい。

以上問題点について述べた。大きく分けて用例数、使い勝手、データ操作の 3 点での問題を確認した。また、今後改良を行ってゆくために、(1) リソースとなるパラレルコーパスの規模を大きくし、(2) 学習者評価を行うことが必須であることを確認した。

本稿の最後に、今回の研究の位置付けについて述べる。

今回の研究では、パラレルコーパスを加工して「用例集」を作成した。その際形態素解析・データの組み替えといった自然言語処理的基礎技術を援用した。従って今回の研究は、言語研究の成果を外国語学習といったニーズに見合った形で社会に還元するために情報工学を援用する、という研究モデルの一事例と位置付けられるだろう。

6．参考文献

6. 1 印刷されたもの

中條清美他　2005　日英パラレルコーパスを活用した英語語彙指導の試み．日本大学生産工学部研究報告Ｂ．2005 年 6 月第 38 巻．17-37．日本大学.

2007　パラレルコーパスを利用した文法発見学習の試み．日本大学生産工学部研究報告 . B, 文系 40, 33-46,

山崎直樹　2009　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」．関西大学外国語教育研究 17, 111-126, 2009-03．関西大学.

Johns, Tim　2000　“Data-Driven Learning: the Perpetual Challenge”, Proceedings of the Fourth Teaching and Learning Corpora（TALC）Conference, Graz, 7/19-23/2000

Bird, Steven　et al.

2009　Natural Language Processing with Python. Oreilly & Associates Inc:

6. 2 Web サイト

Chasen　形態素解析システム茶筌　URL: http://chasen.naist.jp/hiki/ChaSen/

Corpus Concordance English

URL: http://www.lextutor.ca/concordancers/concord_e.html

Juman　日本語形態素解析システム JUMAN 6.0

URL: http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html

Kakasi　KAKASI - 漢字→かな（ローマ字）変換プログラム

URL: http://kakasi.namazu.org/index.html.ja

MeCab　MeCab: Yet Another Part-of-Speech and Morphological Analyzer

URL: http://mecab.sourceforge.net/