句例パラレルコーパスの構築と諸問題(1)

Aufbau des parallelkorpus der Japanisch-Deutschen Beipielphrasen und die probleme

阿部一哉・在間進

Kazuya ABE, Susumu ZAIMA

０．はじめに

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答えることは，外国語教育の根本的な課題の一つであろう。この認識に基づいて阿部（2011）では，実際に日独パラレルコーパスをもとに，日本語の単語から用例を探し出すことのできる用例集を作成し，用例集の作成において援用した形態素解析やデータ組み換え作業といった手法について報告を行った。

現在私たちは，対象をドイツ語学習者からドイツ語を書こうとする日本語母語話者に特化し，ドイツ語を書くときに役立つデータを提供することを目的とした「日独句例パラレルコーパス」の構築を行なっている。

ただし，このコーパスの想定する達成目標は，ドイツ語母語話者のようなドイツ語が書けるようになることではなく，ドイツ語母語話者の言語習慣に可能な限り沿ったドイツ語が書けるようになることである。

本稿では，まず句例パラレルコーパス構築の構想をまとめ，次に具体的な構築の手法について述べる。

１．句例パラレルコーパス構築の構想

１．１．可能な限り大量かつ網羅的な実例収集

このコーパスの構想的特徴の一つ目は，可能な限り大量かつ網羅的に実例を収集し，利用可能にすることである。「可能な限り大量の網羅的実例」を収集するのは，ドイツ語と日本語の対応が多様かつ複雑だからである。その一例として，ドイツ語の動詞 tragen に対する日本語の対応を巻末資料に示す。このような，対応が多様かつ複雑な場合でも，一般的な辞典では，語義と一部の具体例が示されるのみである。

しかし，ドイツ語母語話者の作成したこの種の辞典などに，私たちの書きたいことが記載されていない場合，非母語者の日本人には「正しい」ドイツ語を書く可能性が閉ざされる。したがって，このような状況に対する最も合理的な対処法は，ドイツ語の「正しい」表現，より現実的に言うならば，ドイツ語母語話者が使用した実例を可能な限り大量かつ網羅的に収集し，利用可能にすることである。このようなことも，大量の言語データを処理できる IT 技術が進歩した現在，可能なものになっている。

１．２．使用頻度の導入

構想的特徴の二つ目は，使用頻度の情報を取り入れることである。日本人がドイツ語を書く場合，書きたいことに対して，言語的に「正しい」表現が複数あることがある。たとえば，「生き方を変える」という日本語をドイツ語にする場合，「生き方」を表す Leben は，以下に見るように，動詞 ändern とも verändern とも結びつく。

(1) a.Ich will mein **Leben ändern** - jetzt!

b.Wie kann man sein **Leben verändern**, ...

両者の意味的差異は，たとえば「犬」Hund と「猫」Katze の対立のように，明確な形で意識化できるようなものではない。両者のどちらを使用するのかは，それらの意味的相違を意識しつつも，様々な表現上の観点から判断されると考えられる。

言語的に「正しい」表現が複数ある場合，ドイツ語母語話者によって「よりふさわしい」とされる表現の方を選ぶべきと考えられるが，私たち非母語話者には「よりふさわしい」表現を選ぶ能力はない。したがって，このような場合，最も合理的な対処法は，より多くの母語話者によってより多く使用される表現，すなわち使用頻度の高い表現を選ぶことであろう。ちなみに，大規模コーパス DeReKo に収録されている das Leben ändern と das Leben verändern などの事例数は，以下のようになる。

(2) a.*das Leben ändern*　vs　*das Leben verändern*　　8 対 17

b.*sein Leben ändern*　vs　*sein Leben verändern*　　83 対 16

c.*mein Leben ändern* vs　*mein Leben verändern*　　13 対 6

１．３．句例分析

三つ目は，当面，句例，そしてそれも他動詞の句例を分析の対象にすることである。ドイツ語を書くという目標に置いた場合，第一義的に問題になるのは文である。しかし，たとえば，以下に見るように，日本語とドイツ語の場合，文レベルでは，統語的意味的構造が大きくずれることが多くある。

(3) a.リンゴの木々が白とピンクの花を付けている。

Apfelbäume blühen weiß und rosa.

（＝リンゴの木々が白くそしてピンクに咲いている。）

b.今，交通の至便な場所に住んでいる

（*Ich wohne jetzt, wo es sehr verkehrsgünstig ist.）

Ich wohne jetzt sehr verkehrsgünstig

文レベルのパラレルコーパスを作成する場合，母語話者の協力が常時かつ根幹的な部分で不可欠になるため，作業の推進が不安定になる。たとえ実現不可能と思われても最終的な目標になる文を対象にすべきであるとの考え方もあるであろうが，それと共に，実現可能なところから始めるという考え方もある。私たちは，後者の考え方に立ち，統語的意味的構造の並行性がより多く認めうる句例の分析から始めることにしている。

１．４．IT 技術を用いたパラレルコーパス

四つ目は，IT 技術を用いたパラレルコーパスであることである。パラレルコーパスは，「複数言語について，意味内容がほぼ等しいと考えられる言語表現に対応関係が付いているコーパス」と定義する。

日本語からドイツ語の対応語を調べる従来の和独辞典の場合，用例は，検索した見出し語のところに記載したものしか利用できない。しかし，IT 技術を用いて用例をパラレルコーパス化するならば，検索した見出し語を含む用例をコーパス全体から拾い出すことが可能になる。

また，単に見出し語検索のみならず，複合的な語句を直接入力する検索も可能になる。複合的な語句での検索が可能になるならば，対応するドイツ語の抽出が格段にスピードアップする。さらに言えば，日本語とドイツ語のパラレルコーパスが完成するならば，ドイツ語母語話者が日本語を書く際にも利用可能になる。

なお，日本語とドイツ語のパラレルコーパスとして，広島大学のパラレルコーパス「ドイツ語例文コーパス DJPD（Deutsch-Japanisches Parallelkorpus für Deutschlernende Ver.0.1）がある。

１．５．ドイツ語から分析する根拠，目指すドイツ語

五つ目は，ドイツ語母語話者が使用した実例から句例を収集し，それに日本語を対応付けするという形で行なうことである。日本人がドイツ語を書くときに役立つコーパスを目的とするならば，まずドイツ語にしたい日本語の表現を収集し，それにドイツ語を対応付けするという形が最も自然と言える。しかし，このような場合，またもや，ドイツ語母語話者の協力が常時かつ根幹的な部分で不可欠になる。

それに対して，ドイツ語母語話者が使用した（すなわちすでにネイティブチェックされている）実例から句例を取り出し，日本語を対応付けすることは，日本語の母語話者である私たちには可能である。これが，ドイツ語母語話者が使用した実例から句例を収集し，それに日本語を対応付けするとい

う形で行なうこととした理由である。

なお，この方法論は，以下の２点が問題になる。一つは，このような方法で本当に，日本人がドイツ語を書くときに役立つ句例を収集することができるかどうかということである。しかし，ドイツ語母語話者も日本語母語話者も，この地球という同一空間で現代という同一時代を生きているとするならば，両者の発話内容は，基本的な部分で共通していると十分に仮定できよう。これが，ドイツ語母語話者が使用した実例から句例を収集し，それに日本語を対応付けする上述の方法で十分に行けると判断した理由である。日本人がドイツ語を書くときに役立つパラレルコーパスを構築するという目的は，そもそも程度問題なのである。

もう一つは，日本語的な表現，たとえば「舌を巻く」というようなものが拾い上げられない可能性があるということである。しかし，日本語母語話者には，このような日本語的表現をより中立的な表現，たとえば今述た「舌を巻く」の場合ならば，以下のように「とても驚いた」などに言い換えることができる。

(4) a.　私はその子の見事な歌いっぷりに舌を巻いた。

→ …とても驚いた／→ …驚きあまり言葉を失って聞いていた。

b.　Ich war sehr erstaunt, wie gut das Kind sang.

Sprachlos vor Bewunderung hörte ich das Kind singen.

＊ドイツ語例は『アクセス和独辞典』（三修社）より

このような言い換えの能力を前提とすることは，本研究の目的が，ドイツ語母語話者のようなドイツ語が書けるようになることではなく，ドイツ語母語話者の言語習慣に可能な限り沿ったドイツ語が書けるようになることであるとするならば，上述の問題点は必ずしも本コーパスにとって致命的なものにならないであろう。

２．句例パラレルコーパスの構築手法

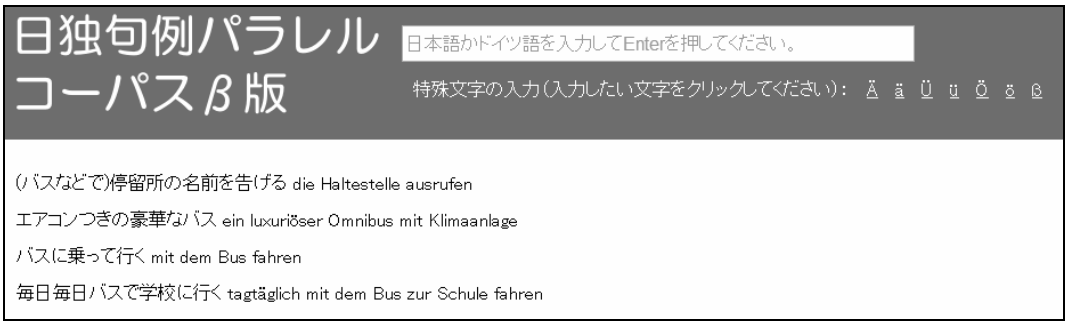
前節では，句例パラレルコーパス構築の構想を述べた。本節では，パラレルコーパス構築の具体的な手法について考察を行う。考察にあたって，既存の句例データ [(2)] に基づき，試行的にパラレルコーパスの構築を行い，「日独句例パラレルコーパスβ版」として，以下の URL にアップロードした。

(5)　日独句例パラレルコーパスβ版

http://atomilang.atomi.ac.jp/parac/

上の URL をブラウザで開くと，（6）に図示する Web ページが立ち上がる。

(6)



対訳表現を知りたい表現を「日本語かドイツ語を入力して Enter を押してください。」欄に入力すると，ヒットした事例が下欄に表示される。なお，図に示されているデータは，「バス」という検索文字列に対して返された検索結果である。

このように，使用法としては非常に単純なものを想定しているが，すでに述べたように可能な限り網羅的な事例を用意して，日本語ドイツ語両方から引ける検索機能を備えていることで，有用なツールになると考えられる。

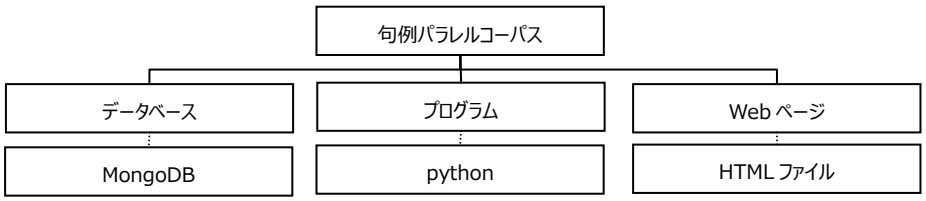
以下句例パラレルコーパスの構築手法として，２．１．節では句例パラレルコーパスのプログラムとしての全体的な構造，２．２．節では，句例パラレルコーパスの最も重要な機能である検索機能について見る。そして２．３．節では，句例のデータ構造と処理の流れについてまとめる。

２．１．句例パラレルコーパスの構造

本節では，句例パラレルコーパスのシステムとしての全体的な構造について見る。上で見たような仕組みの裏側では，使用者の要求に応じて，データベースに格納された句例データを選別し，Web ページ上に表示する，という一連の処理が行われている。そのためには単純な HTML ページを記述するだけでは不十分で，膨大な句例データを管理するデータベースと，検索などの必要な処理を記述するプログラミングが必要になる。句例パラレルコーパスでは，データベースとして MongoDB [(3)]，プログラムを記述するためにプログラミング言語 python をそれぞれ使用している。

なお，このように，インターネット上で一般の使用者に対してサービスを提供するプログラムを総称して Web アプリケーションという。句例パラレルコーパスの Web アプリケーションとしての構造を図示すると以下のようになる。

(7)



Web アプリケーションを構築するためには，必要な機能をあらかじめ用意した「フレームワーク」が数多く開発されている。開発者はこれらのフレームワークを利用することで，比較的手軽に Web アプリケーションの開発を行うことができる。私たちはそのうちの一つである Flask を利用した。

この句例パラレルコーパスにおいて，最も重要なのが検索の仕組みである。第 1 節で確かめたような，日本語・ドイツ語の両方から引けて，検索語ひとつ，あるいは複数の検索語の組み合わせが現れるすべての用例を抽出するような検索機能を実現するためには，いくつかの解決するべき問題点がある。そこで次の２．２．節では，句例パラレルコーパスにおける検索方法の実装について見る。

２．２．句例パラレルコーパスにおける検索方法

本節では，句例パラレルコーパスにおける検索方法について見る。検索語が含まれる句例全てを抽出するような検索を行うためには，以下に示すように，予め句例で用いられている語を使って索引を作っておく必要がある。

(8)　**句例**：バスで学校に行く　　**索引語**：「バス」「で」「学校」「に」「行く」

実際の検索を，句例そのものでなく，索引語に対して行うことで検索語が含まれる句例を抽出する仕組みを実現することができる。

(8) の「バスで学校に行く」の場合，「バス」「で」「学校」「に」「行く」という検索語いずれを入力しても，同じ句例を返すことになる。

また，複数の検索語の組み合わせからの検索，例えば「バス」＋「で」＋「行く」という組み合わせからの検索を可能にすることで，「バスで行く」という検索文字列を使っても「バスで学校に行く」という句例をヒットさせることが可能である。

このように，検索対象である句例に対して，予め索引語を付与しておくことによって，柔軟な検索が可能になる。

もっとも，膨大な句例データひとつひとつに，上で見たような検索語を手作業で付与することは，ほぼ不可能である。そこで本研究では，日本語には形態素解析エンジン MeCab を，ドイツ語には treetagger をそれぞれ使用し，索引語を付与している。

形態素解析とは，テキストを語に切り分け，個々の語の品詞情報などを解析する処理のことである。

例えば，上例「バスで学校に行く」を，MeCab で処理すると以下の結果が返される。

(9)
```
バス　名詞,一般,*,*,*,*,バス,バス,バス
で　助詞,格助詞,一般,*,*,*,で,デ,デ
学校　名詞,一般,*,*,*,*,学校,ガッコウ,ガッコー
に　助詞,格助詞,一般,*,*,*,に,ニ,ニ
行く　動詞,自立,*,*,五段・カ行促音便,基本形,行く,イク,イク
EOS
```

MeCab は，「バスで学校に行く」という表現を，各行左端のような形で単語に切り分ける。そして各行右側のように，語の情報をコンマ区切りで返してくれる。MeCab の場合，より詳細には，以下の情報を返してくれる。（例は「行く」を解析した場合の出力）(4)

(10)

表層形	品詞	品詞細分類 1	〃 2	〃 3	活用形	活用型	原形	読み	発音
行く	動詞	自立	*	*	五段・カ行促音便	基本形	行く	イク	イク

句例を語に分割する作業は，このように形態素解析プログラムを利用することで，自動的におこなうことができる。

なお，実際にプログラムを記述していく中で，語の変化形（活用形）に検索語をどのようにマッチングさせるかということが問題になった。例えば「秘密をしゃべらないでおく」の索引語（助詞を除いてある）は，単純な語の切り分けに基づいた場合，「秘密」「を」「しゃべら」「ない」「で」「おく」になる。

しかし，実際の使用を想定した場合，「しゃべる」という検索語に対しても，「しゃべら」がヒットすることがよりよいだろう。そこで句例パラレルコーパスでは，切り分けられた語の，表層形ではなく上表で言うところの「原形」に基づいて索引語付与を行なっている。

なお，索引語が原形に基づく場合，検索文字列に語の変化形が含まれると逆にヒットするべき句例がヒットしなくなってしまうので，検索の都度，検索文字列にも形態素解析を行い，原形の語の集合に変換してから，マッチングを行うようにしている。

句例に索引語を付与する際に考慮に入れるべきもう一つの点は，日本語は助詞，ドイツ語は冠詞のように，機能語は経験的に見て頻度が多いため，索引語として取り出すことはあまり意味が無いように思われる。

上で，句例パラレルコーパスでは，検索の都度，検索文字列に対しても形態素解析を行なっていると述べた。その際，上表の「品詞細分類」情報を利用し，「助詞」と判断されている語を検索語の集合から除外するようにしている。従って，句例「学校に行く」という検索文字列は「学校」「行く」という二つの検索語の集合に変換される (5)。

２．３．句例のデータ構造と処理の流れ

本節では，句例のデータ構造と検索の一連の処理の流れについてまとめる。句例パラレルコーパスで利用している MongoDB では，句例は以下の形で保存されている。ここでは「緻密なやり方で auf subtile Art und Weise」を例に取った。

(11)

行数	データ	ブロック	
1	{		データ全体
2	"_id"：ObjectId("50cdfff6649d153f900001f5"),		
3	"jsent"：{	日本語データ	
4	"lemma"： ["緻密", "だ", "やり方","で"],		
5	"word"： ["緻密", "な", "やり方", "で"],		
6	"pos"： ["名詞","助動詞","名詞","助詞"]		
7	},		
8	"dsent"：{	ドイツ語データ	
9	"lemma"： ["auf", "subtil", "Art", "und", "Weise"],		
10	"word"： ["auf", "subtile", "Art", "und", "Weise"],		
11	"pos"： ["APPR", "ADJA", "NN", "KON", "NN"]		
12	}		
13	}		

2 行目は，データベースが句例データに対して自動的に付与する一意の識別番号である。3 行目以降データ全体は大きく分けて日本語データ（3～7 行目）と，ドイツ語データ（8～12 行目）からなる。

日本語データの場合，4 行目に句例を構成する語の原形の集合が，5 行目に表層形の集合が，6 行目に品詞の集合が，それぞれ句例の語順通り配列されている。

ドイツ語データの場合，9 行目に句例を構成する語の原形の集合が，10 行目に表層形の集合が，11 行目に品詞の集合が，それぞれ句例の語順通り配列されている。

このデータに対するプログラムにおける処理の流れは，以下の通りである。

（1） 検索文字列を Web ページから受け取る。

（2） 検索文字列を，検索語の集合に変換する。

すでに述べたように，検索文字列を検索語の集合に変換する際，助詞などの機能語は除外される。例えば「学校に行く」は『「学校」，「行く」』という検索語の集合に変換される。

（3） 検索語の集合が，句例データの原形の集合に含まれるか否かマッチングを行う。

例えば，『「学校」，「行く」』に対して，句例「毎日毎日バスで学校に行く」はヒットするが句例「バスに乗って行く」はヒットしない。

（4） ヒットした句例データの表層形の集合を使って，日本語，ドイツ語それぞれについて句例文字列を合成する。

日本語は，上表の表層形「"緻密", "な", "やり方", "で"」を，引用符を除いて結合して日本語

の句例文字列をつくる。ドイツ語は，上表の表層形 "auf", "subtile", "Art", "und", "Weise" を，引用符を除き半角スペースで結合して，ドイツ語の句例文字列をつくる。そのようにして得られた，日本語とドイツ語の句例文字列をタブ記号を使って結合し最終的な句例文字列として利用している。

（５） 句例文字列の集合を Web ページに表示する。

（１）（５）のような，汎用性の高い処理は，Web アプリケーションフレームワーク Flask の機能を利用している。

以上，本節では句例パラレルコーパスにおける句例データの構造と，このデータに対する一連の処理について見た。

３．おわりに

本稿では，第 1 節で日独句例パラレルコーパス構築の構想について述べ，第 2 節で構築の手法について述べた。

おわりに，本節では今後の課題として，句例データの収集と，Web アプリケーションの機能拡充について述べる。

句例データの収集方法については，第一節で方向性をまとめたが，実際の作業としては様々な収集方法を試行的に行なっている段階である。例えば，在間・カン（2011）の成果を援用し，他動詞の zu 不定詞句の事例にもとづき，コンコーダンサを使って目的語を抽出する手法を考案した [(6)]。言語使用データとしては，主としてベルリン・自由大学で開発保守がなされている COW（Corpora from the Web）を利用し，作業を進めているところである。

Web アプリケーションの機能拡充については，検索方法と，検索結果の表示方法について改良を行なっていく必要がある。そのために，他の類似する Web アプリケーションの仕様を参考にする，使用者評価を実施するなどの第三者的視点からの検証が不可欠である。

【参考文献】

阿部一哉 (2011) 日独パラレルコーパスを利用した用例集の作成. 跡見学園女子大学文学部紀要 46, A. S. 77-92.

在間　進，カン・ミンギョン（2011）コーパスに基づくドイツ語文形成規則の分析 − 主に方法論的考察について−．ドイツ文法理論研究会編．ENERGEIA 第 36 号．S.59-73.

Banker, Kyle（2012）MongoDB イン・アクション．Sky 株式会社 玉川竜司訳．オライリー・ジャパン（原著：MongoDB in Action. Manning: Shelter Island）

【Web サイト】

ドイツ語例文コーパス DJPD　http://www.vu.hiroshima-u.ac.jp/deutsch/

COW – Corpora from the web　http://hpsg.fu-berlin.de/cow/?action=home&lang=de-DE

Das Deutsche Referenzkorpus – DeReKo　http://www.ids-mannheim.de/kl/projekte/korpora/

Flask (A python Microframework)　http://flask.pocoo.org/

MeCab: Yet Another Part-of-Speech and Morphological Analyzer
http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

Python Programming Language – Official Website　http://www.python.org/

TreeTagger - a language independent part-of-speech tagger
http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

【巻末資料】ドイツ語動詞 tragen と共起語と日本語の対応語

1) Jacke, Kleid, Uniform　＝着ている（上着／ワンピース，ドレス／制服）
2) Rock, Strümpfe, Schuhe　＝はいている（スカート／靴下／靴）
3) Handschuhe, Ring, Uhr　＝はめている（手袋／指輪／時計）
4) Schmuck, Kette, Perücke,　＝付けている（アクセサリー／ネックレス／かつら
Orden, Parfüm, Ohrhörer, Hörgeräte　／勲章／香水／イアホン／補聴器）
Blume（... im Kopf tragen）　＝…に付けている／挿している *（花；…を頭に）*
5) Hut　＝かぶっている（帽子）
6) Bart　＝はやしている（ひげ）
7) Brille　＝かけている（眼鏡）
8) Waffe　＝身につけている，携えている（武器）
9) Haare　＝…している（髪；lange Haare tragen）

注

(1) 本稿は，日本独文学会 秋季研究発表会（2012 年 10 月 13 日（土）・10 月 14 日（日））口頭発表（語学）における，阿部，在間による発表「句例パラレルコーパス構築とその諸問題」を元に加筆・修正を行ったものである。

(2) 既存のデータとして，東京外国語大学在間研究室で研究用に作成した事例集を使用した。事例集自体は非公開である。

(3) MongoDB は，python のディクショナリ型にほぼ完全にマッチング可能な bson 形式でデータを管理しているため，データベースのデータ構造とプログラミング言語のデータ構造の差異から生じるインピーダンス・ミスマッチ問題を比較的容易に回避できる。

(4) http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

(5) 今回実装した，検索方式は全文検索と呼ばれるが，書くための句例コーパスという性質上，本来は完全一致検索が望ましい。完全一致検索でも，ニーズに合った検索結果が得られるよう，基盤となる句例データの質的量的な拡充を今後の課題としたい。

(6) この内容については，阿部，在間で報告を行った。注 1 参照。