言語研究対象としての中国語Wikipediaの可能性

氷　野　善　寛

Chinese Language Wikipedia as a Potential Linguistic Object

HINO Yoshihiro

One issue is what type of language-related materials needs to be collected when constructing a corpus that targets today's Chinese language and when performing a consideration of the Chinese language. And, because copyright protection is required for a great deal of language-related material, attention needs to be taken as well when collecting and using modern materials. Under such circumstances, the Internet has recently been chosen for use as a huge language database, with research conducted using existing search and information retrieval tools. In this case, however, the results vary each time and reproducibility of the research thus becomes a problem. I therefore paid attention specifically to the Chinese version of Wikipedia as the language-related material for Chinese language research. In this paper, I considered the specific characteristics of Wikipedia, noted its possibilities as language-related material, and added the experimental consideration of using Wikipedia as a corpus.

Keyword：Wikipedia, Chinese Corpus, Digital archives, Database

はじめに

データベースの重要性は様々な分野で説かれ、研究対象・研究分野によってそれぞれのニーズに応じたデータベース構築が模索されている。ただ、近年では、データベースという言葉が先行し、何でもかんでもデータベースにしてしまえば良いと思われている節がある。しかし、ある研究分野に対して必要なデータベースを、思ったとおりに構築できているのかどうか、という点についてはまだまだ検討の余地があるのではないだろうか。

文献資料のディジタル化について見れば、最近では国内外を問わず、書籍のディジタル化及びアーカイブ化作業を行っているところが多々ある[1)]。たしかにネットワーク上で図書館の貴重書にアクセスできることは、一昔前に比べると研究者にとっては非常に便利になったと言えるだろう。ただ、データベ

1）国立国会図書館の「近代デジタルライブラリー」（http://kindai.ndl.go.jp/index.html）、「関西大学図書館特別蔵書」（http://www.kansai-u.ac.jp/library/library/）など、図書資料の電子化し無料で公開する例が多い。

ースが構築されたからといって、それらが研究者にとって本当に有意義なものとなるかどうかは別問題である。たとえば文献研究を行うものには、書籍の中にある文字データよりもその書籍が持つ外的な情報、紙質、色合いといった情報をより重視することがある。文献をディジタル化して、ネットワーク上にそのデータを供給する場合、「掲載する」ことに重点が置かれ、データを圧縮してしまいがちである、その結果として、画像品質に重点を置かず、これらの属性情報を再現できない状態でオンライン上に掲載するといった例がよく見られる。結果として、せっかくディジタル化された貴重な資料でも、これらの文献研究を行うものにとっては無用のものとなってしまう。だからといって、そのデータベースは無用であると即座に判断することはできないが、文献をディジタル化し、それを給する場合、誰を利用者として考えているのか、需要と供給を一致させるにはどのようにデータベースを構築するべきか、構築する前の段階でしっかりと議論する必要がある。

一　言語研究にとってのデータベースとWikipediaの持つ可能性

これまで筆者は、言語研究に関わるデータベースの構築に携わってきた。「関西大学現代中国語コーパス」[2]と「近代漢語文献データベース」[3]である。これらのデータベースの構築及びテキストの整理に関しては、中国語の形成について言語的観点から研究を行う、という点を重視している[4]。そのためどちらのデータベースについても語彙検索を重視した設計を行っている。ただ、それぞれ用途と研究対象とするものが異なる。「現代中国語コーパス」では、新聞、小説や中国語教材を中心に、現代中国語に関する言語資料を収集し、テキストデータベースを構築している。「近代漢語文献データベース」では、清朝期に西洋人や日本人が利用した中国語教材、聖書やイソップなど、中国に西洋の知識を伝えるために用いられた翻訳書を中心にディジタル化を行っている[5]。前者については現代語を対象としているため、著作権保護などの問題がある。後者については、19世紀の書籍が大半を占めるため著作権については問題にはならない。このデータベースについては、文献研究と言語研究の両方に用いられるよう、文献画像の表示と語彙検索といった特徴を持たせられるよう整理を進めている。

このように、言語接触研究及び教材研究において、データベースを用いての研究を考慮し、これまでこの二種類のデータベースを構築し、運営してきた。今回、中国語の研究を「文化交渉」という立場に立脚し、言語的なアプローチを加えていくためには、これらのデータベースをどのように発展させ、利用する価値のあるデータベースへカスタマイズしていくかを考えていく必要がある。そこで、今回はま

2)「現代中国語コーパス」（http://china.fl.kansai-u.ac.jp/）

3)「近代漢語文献データベース」（http://www.csac.kansai-u.ac.jp/db.html）

4) これまでも『四庫全書』や『中国基本古籍庫』のような中国の文献資料を対象としたディジタル化作業は多く行われている。これらの中国の文献資料を扱うデータベースが文献研究の立場に立脚したものであったのに対して、「近代漢語文献データベース」では、文献資料的研究に用いるツールだけではなく、語彙研究を行えるような機能を追加している。

5) この種の書籍は『四庫全書』や『中国基本古籍庫』など、いわゆる中国の古典の大規模なディジタル化の範疇に入っておらず、ディジタル化されていない。ただし当時の中国語の諸相を知るためには有用な資料となりうる。

ず「現代中国語コーパス」の言語資料について考えてみたい。

コーパスとは、ディジタル化された自然言語の文書から成る巨大なテキストデータのデータベースである[6)]。近年、言語研究の分野においてもデータベースの重要性は説かれ、コーパスを用いた研究などが盛んに行われている。中でも現代語を対象としたコーパスの構築を考える際、問題となるのが、言語資料の収集対象であり、収集方法である。さらに、テキストの著作権とその運用に関する問題が絡んでくる。また、英語や日本語については近年大規模な言語資料の収集が行われ整備が進められている。英語圏では大規模なコーパスが構築され、言語研究に大きく寄与をしているだけではなく、近年では研究だけではなく、学習にも用いられるようになっている。日本語においても国語国立研究所を中心に、明治から現代にいたる日本語の全貌を把握するために、言語コーパスを構築するという「KOTONOHA」計画を実施するなど、大規模なコーパスが試験的に構築されている[7)]。

それでは筆者がフィールドとする中国語ではどのような状況かと考えると、たとえば北京大学漢語語学研究中心の「CCL語料庫」[8)] で人民日報や小説などのデータを用いたコーパスを構築しているほか、香港城市大学の「LIVAC共時語料庫」[9)] が行っている中国国内の数地点の新聞を収集したコーパス、台湾中央研究院が構築する「現代漢語平衡語料庫」[10)]、関西大学の「現代中国語コーパス」などを例がある。どれも研究利用を目的としてテキストデータの収集を行っているが、多くのデータベースが新聞のテキストデータを扱っており、必ずしも誰もが利用できる幅広い資料を扱ったオープン性の高いコーパスが構築されているとは言い難い。

このような状況において、インターネットを巨大な言語データベースとみなし、Googleのような検索ツールを用いて研究利用することを主張している[11)]。インターネット上のソースを利用することは、オープン性の高い整備された中国語コーパスが限られる中、自由に使えるデータベースということで研究に利用することができる。とりわけ、辞書に掲載されていない新語の研究を行う場合には、有用であることが多い。ただし、その反面、インターネット上の情報は日々変化し、検索の規則や条件も同様に変化する可能性があり、そのため、同様の条件下で検索を行ったとしても、必ずしも同じ結果を得られるとはかぎらないなど、研究の再現性が難しいといった点が問題となる。

そもそも、中国語研究にとって、そもそも英語や日本語と同じようにコーパス研究というものが有用であるのかといった問題もある。この問題点が、中国語コーパスの大規模な利用につながらない現状を作り出していると考えることができる。

以上の点を整理すると、オープン性の高い言語資料を対象としてデータベースを構築する際、

6）オンライン上で公開されているコーパスとは別に、個人でテキストデータで収集して、エディタなどを用いてコーパス利用するケースもあるが、インターネットの普及によりネットワーク上で大規模なコーパスが構築することは、研究者だけではなく学習者の利用をも促すこととなり、幅広い面での利用が想定できる。

7）国立国語研究所言語コーパス整備計画KOTONOHA（http://www2.kokken.go.jp/kotonoha/）

8）北京大学漢語語学研究中心「CCL語料庫」（http://ccl.pku.edu.cn）

9）香港城市大学「LIVAC共時語料庫」（http://rlcondor.cityu.edu.hk/）

10）台湾中央研究院「現代漢語平衡語料庫」（http://www.sinica.edu.tw/SinicaCorpus/）

11）氷野善寛　2007「关于汉语语料库」『中日研究生国际论坛 2007 汉语汉文化论丛』（白帝社）p39～50

1．テキストデータの著作権
2．検索対象となるデータの再現性
3．中国語を対象とした検索方法

の3点が問題点となる。そこで第1の問題については、テキストの幅広い改変と二次使用等を認めているWikipediaに着目した。第2の問題については、ある種のデータ群を用いて、再現性のある研究を行う場合には、検索対象となるテキストをひとところに集積し、それに対して検索処理を行う必要がある。Wikipediaを対象として検索処理を行う場合、Wikipediaのデータは定期的に全データのバックアップデータを圧縮したものを配布されていることから、このデータを利用することができる。また検索ツールとしては、関西大学の「現代中国語コーパス」を用いれば、検索を再現することも可能である。第3の問題については、Wikipediaをインターネット上のサイトで利用した場合、主な検索対象は見出し語とであり、中身の検索についてはそれほど詳細な検索ができるわけではなく、Google検索を組み合わせることや検索しなければならず、検索効率については決して高くない。しかし、この点については、関西大学の「現代中国語コーパス」に搭載した検索ツールを用いることで、対応することができる。

このような理由から言語資料として「Wikipedia」中国語版に注目したが、私の研究の出発点は方言と標準語の接触による、新語の形成についてであり、これまで、言語接触、言語交流による地域標準語の形成と、それが全国で標準語として通用するまでのメカニズムについての研究を行っていたが、この点からもWikipediaというものは興味深い研究対象となる。

Wikipediaというのは、インターネットという非常に限られた空間に存在し、多種多様な人が参加しているため、言語的にも、不安定かつ、未知の要素が含まれている。そのため扱いにくい素材であることは間違いない。ただ、この事が一般的な環境と異なり、はやい速度でその中で使われている言葉に影響を与えることも想像できる。そこで、中国語Wikipediaがいったいどういうものであるのかについて見直し、その利点について考察していく。

二　Wikipediaについて

Wikipedia[12] はインターネットを介して、誰もが閲覧・執筆・編集できる百科事典である。2001年1月に英語版の作成から始まった同プロジェクトは2008年1月の時点で200以上の言語で作成されている。中国語版[13] は2002年10月に創設され、2007年11月の時点で、16万項目の記事があり、51万ページ存在する。単純に各ページに1000字程度で構成されていたと考えても5億字のデータベースということになり、その規模の大きさがうかがえる。

Wikipediaの特徴を考える場合、ひとつの記事に対して、多数の執筆者が「共同編集」という形で記事を作り上げていっていることにある。これは一つの見出し語に対して、一人の執筆者が書くのではなく、不特定の多くの執筆者が書くことにより、精度の高い記事を生み出していくものであると考えられ

12)「wikipedia日本語版」（http://ja.wikipedia.org/）
13)「wikipedia中国語版」（http://zh.wikipedia.org/）

る。また一般的に辞書は出版された段階ですでに過去の記事であるのに対して、Wikipediaで作成される記事については完成という概念がなく、常に書き換えられていく可能性も含んでいる。これらの点は、記事に言語的な多様性を生み出せるだけでなく、どのように書き換えられていくかを考察できるなど、興味は尽きない。

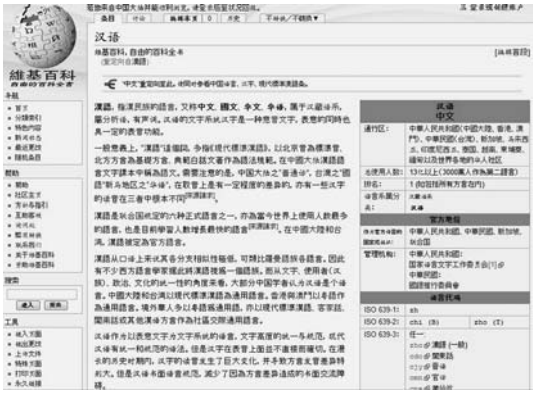

図１．Wikipedia中国語版

次に、このような過程で作成されたWikipediaの文章は、一般的な文章と著作権の扱いが異なる。一般的にある人が書いたものをまた別の人が編集するという行為は、著作権の禁じるところである。そのためWikipediaでは、GNU Free Documentation License（GFDL）[14] に基づいて文章の利用を許諾し、文章の無断での改変、再配布、二次利用を許可し、執筆者にそのことに同意した上で、記事を作成、利用するよう求めている。そのため、当然ながら、Wikipediaでは他人が著作権を保有する記事を無断で掲載することを禁じている。このように、Wikipediaの最大の特徴は、技術的には共同編集を可能とした「Wiki」[15] を、著作権的には利用者にGFDLの規定に準拠させたことで、他者の文章を書いたものを自由に編集することを容易にしている。

また、Wikipediaにはこういった辞書プロジェクト以外にも「Wikisource」[16] というプロジェクトも存在する。同プロジェクトでは、あらゆる言語の公文書、歴史書、著作権が既に消失したテキストの収集が行われるとともに、校閲作業が行われている。これらのテキストについても、同様の権利で構築されている。

三　中国語Wikipediaの特徴と言語接触の場としてのWikipedia

Wikipedia日本語版の記事については、メインカルチャーよりサブカルチャーのほうが記事の内容が多いという指摘がある[17]。中国語版Wikipediaについては、科学技術方面の用語については英語からの翻訳が多く、サブカルチャーについては日本語版からの翻訳が多いるように見受けられる。Wikipedia自体は見出し語があり、それに対する記事が書かれている。そして、その記事の編集内容について「討論」する「場」があり、それぞれ文体が異なる。このように英語及び日本語からの翻訳としての「場」と翻訳に依存しないオリジナルの記事、討論の場のように、Wikipediaには３つの場が存在し、データを豊富なものとしている。

さらに言語的な問題を見ると、中国語版は他言語のWikipediaと比べると特異な点がある。まず、他

14）http://www.gnu.org/licenses/fdl.html

15）Wikiとはインターネット上でテキストを書き換えるシステムの一種であり、このシステムは、ネット上における共同編集やWebページの構築に用いられることがある。

16）http://zh.wikisource.org/

17）藤原敦　2006「Wikipediaとは何か」『漢字文献情報処理研究』

のWikipediaが少なからず国家という枠組みにしばられているのに対し、漢語（普通話・国語・華語）という漢字文化圏を一元的なものとみなし、サイト構築作業を行っている点である。中国語には簡体字と繁体字の区別があるため、中国語Wikipediaには漢字処理の問題がある。簡体字を扱うのは中国大陸、シンガポール、マレーシア、繁体字を扱うのは香港、台湾である。同一の漢字であることには違いないが、同じ漢字を使っていても、単語レベルや文レベルで異なるものもある。とりわけ、新語を中心に異なる箇所がある。このため、Wikipedia中国語版では当初、簡体字版と繁体字版で合計4種類のサイトが設置されていた。そのため、記事もたとえ同じ項目であっても記事はそれぞれの地域で作成されていたが、2005年ごろから利用者の設定により、すべてを一つのフィールドで扱い、表示する漢字を切り替えられるようになった。ただし、それでも使用言語は地域によって大きな差が生じることがある。たとえば、「プリンター」という単語をWikipedia中国語版で確認してみる、

［簡体版］

【打印机】

打印机是一种电脑输出设备，可以电脑内储存的数据按照文字或图形的方式永久的输出到纸张或者透明胶片上。…（以下省略）

［港澳繁体版］

【打印機】

打印機是一種電腦輸出設備，可以電腦內儲存的數據按照文字或圖形的方式永久的輸出到紙張或者透明膠片上。…（以下省略）

［繁体版］

【印表機】

印表機是一種電腦輸出設備，可以電腦內儲存的數據按照文字或圖形的方式永久的輸出到紙張或者透明膠片上。…（以下省略）

（下線は筆者による）

のように変換される。簡体字と繁体字の間だけではなく、繁体字同士でも見出し語の表示か変わることもある。このように表示する言語により見出し語が自動的に変わることがわかる。さらに簡体字と繁体字どちらにも変換されていない状態で表示すると、次のようにひとつの文章の中に、繁体字と簡体字が入り混じっていることが分かる。

［変換無］

【打印机】

打印机是一种電腦输出设备，可以電腦内储存的数据按照文字或图形的方式永久的输出到纸张或者透明胶片上。…（以下省略）

（下線は筆者による）

このことはすなわち、少なくともこの文章を編集する際、異なる地域のユーザーが編集作業を行っていることが分かる。Wikipedia上での語彙レベルでは、大きな異なりがある語彙の場合、自動的に変更されるシステムを採用しているが、文レベルになった場合、その変化は見られない。そのため、異なる背景を持つ言語が接触した場合、各地域の言語にどのような影響が生じるのか、調査する価値はある。このことは、これまでには見られない本人が「意識をしない」状態での「語彙接触」である。特にWikipedia上では、このように編集ユーザー気が付かない間にマクロレベルの語彙接触が行われている可能性があり、このことがWikipediaの語彙、文体を研究する上で、非常に興味深い点となることを指摘することができる。

四　Wikipedia研究の可能性と現代中国語コーパス

本来Wikipediaは百科事典であるため、ネット上に展開されたWikipedia自体を言語コーパスとして利用できないことは先に述べた。特に言語的研究アプローチを加える場合、利用するコーパスでは、単なる語彙検索ではなく、詳細な検索を行う必要がある。そのため、Wikipediaから定期的に出されるバックアップデータのアーカイブに注目し、そのデータを関西大学の「現代中国語コーパス」に組み込みこむことで、コーパスとしての利用と、研究の再現性に関する問題を解決できないか検討することとする。たしかにWikipediaを言語資料としてコーパス化することについては批判が多いのも事実であるが、用途とテキストの種類さえ理解していれば、この言語資料のデータは大いに価値のあるものへと変わる可能性は十分にある。

中国語版Wikipediaのデータについては、定期的にアーカイブを公開しており[18]、半年に一回程度の割合で公開されている。そのファイルから記事部分と討論部分のテキストファイルを抜き出し整理し、「現代中国語コーパス」に読み込むことで、コーパスとしてWikipediaを利用することは可能であると推察できる。日本語研究では長谷部陽一郎氏がWikipediaの日本語コーパス化を試みており、実際にその圧縮ファイルからデータ抜き出しの方法については、氏が開発したツールとともに論文にて紹介している[19]。この論文で紹介されている方法と氏が作成したツールは日本語Wikipediaに特化したものであるが、これを中国語に応用することは原理上可能であると考え、中国語Wikipediaのバックアップデータから中国語データの抜き出しを試験的に行った。さらにこのファイルを関西大学の「現代中国語コーパス」を設置した。その結果は、ファイルの抽出方法や読み込み方法にまだまだ問題はあるもののデータの抜き出し、データベースとして設置するのは問題ないことが分かった。

ここで設置した「現代漢語コーパス」は、2003年末に関西大学に設置、2007年にプログラムをリニューアルした簡体字中国語に対応したコーパスで、北京南宸電子技術有限公司と関西大学でプログラム開

18）http://download.wikimedia.org/zhwiki/

19）長谷部陽一郎　2006「Wikipedia日本語版をコーパスとして用いた言語研究の手法」『言語文化』9(2)

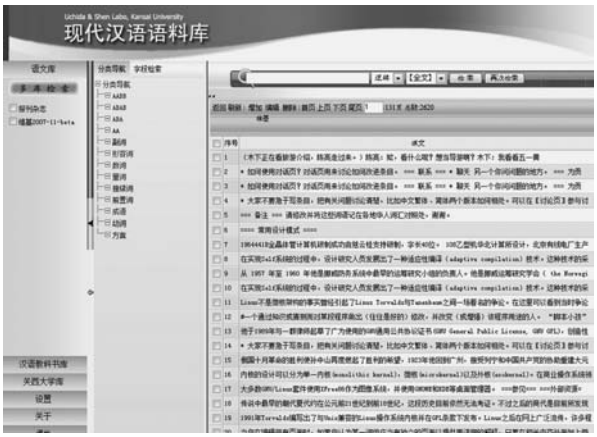

図２．Wikipedia中国語版の一部を「現代中国語コーパス」に組み込む

発を行い、関西大学のサーバーで運用公開している。このコーパスには人民日報数年分及び著作権の切れた中国語テキストなどのデータを入れている。このコーパスに導入するメリットとしては、テキストを登録する段階で、登録する文章をインデックス処理したうえで、別に準備した辞書群と照合させさらにデータを整備することで、検索効率を飛躍的に高めている点である。これにより大容量のテキストデータの検索に耐えること、ナビゲーション機能や複合検索を利用することにより高度な検索処理ができ

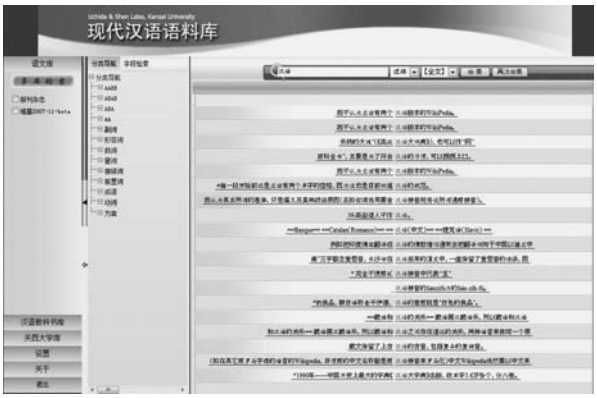

図３．「関西大学現代中国語コーパス」：一般的な語彙検索

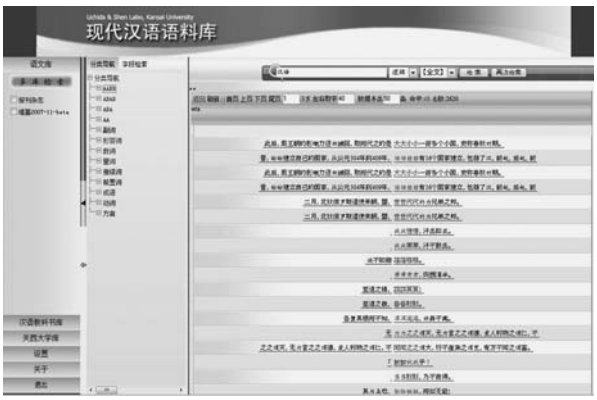

図４．「関西大学現代中国語コーパス」：ナビゲーション機能を用いた検索

る。Wikipedia内部に設置された検索ツールの検索効率に比べると言語研究においては有利である。

おわりに

本論では、言語研究におけるデータベースを構築という観点から、Wikipediaを「現代中国語コーパス」に組み込むことの意義と、Wikipedia自体を研究対象とすることの意義について述べた。まだまだ準備段階で、実際にどのような結果が出てくるかは未知数ではある。しかし、Wikipediaは単なるインターネット上の辞書というだけではなく、多種多様な人が参加し、常に編集が行われていることから、これまでとは全く異なる枠組みで言語接触が起こる可能性がある「場」であるという一面があるということがわかった。また言語資料としてコーパスに組み込むことも、Wikipediaの持つ性格から考えて価値があり、今後これを考察することは、現代中国語の言語形成の在り方の一端を考えるうえで価値があると考えられる。