シンポジウム報告

総論--海外におけるテキスト・データベース開発と著作権

長瀬 真理†

適切なご紹介をしていただきましたので、少々気が楽になりました。きょうは、海外のテキスト・データベースの著作権ということを中心に私はお話させていただきます。簡単なメモ書き風のレジュメを1枚と、それから資料1枚を用意いたしましたので、大体それに沿って話を進めたいと思います。

最初に名和先生がおっしゃってくださいましたが、今回は、文学や古典のテキスト・データベースを中心にお話したいと思っております。"テキスト・データベース"という言葉について聞きなれない方もいらっしゃるかと思いますが、図書館などの目録を入力した書誌データベースと区別いたしまして、テキストの全体を丸ごとコンピュータに入力したものを、一応テキスト・データベースと呼んでおります。英語では、「マシーン・リーダブル・テキスト」とか、あるいは、「エレクトロニック・テキスト」と呼ばれておりまして、直訳すれば「機械可読テキスト」あるいは「電子化テキスト」といわれています。一般に検索をしたり、それから修正、加工など、コンピュータ処理が出来るようなフルテキストをテキスト・データベースと呼んでいるとお考えいただければよいかと思います。

それからもう1つ、これも、いま先生がおっしゃってくださいましたが、今回はテキスト・データベースの中でも文学とか古典を中心に考えております。社会科学の、たとえば法令のデータベースとか議事録に関しましてはすでにいろいろな所で出回っており利用されているということなので、今回は文学・古典といったものを中心にお話したいと思います。

私のデータベースとのかかわりは、まずはユーザーサイドからでございまして、実は専門がギリシャ哲学、プラトンを研究しておりまして、約8メガバイトぐらいのテキストをアメリカのカリフォルニア大学から購入いたしまして、それで文体研究を始めたのが最初です。そのあと、ひょんなことで『源氏物語』、これも小学館のバージョンですが、それのデータベースと、それから、サイデンスティッカーの英訳本、『The Tale of Genji』の両方のデータベースを作成いたしました。

そして、残念なのですが、日本国内には海外の研究者にサービスをする機関がございませんので、オックスフォード大学のコンピュティング・サービスという所に委託をいたしまして、公開をしております。というわけで、一応ユーザーサイドと、それから制作というか、供給の両方の経験を持っているということです。

とりあえずは、先に先生がシンポジウムの経緯について簡単にお話しくださいましたのでこのところは省きますけれども、日本では非常にテキスト・データベースの開発が遅れております。欧米では1960年代からもうすでに、古典、あるいは文学のテキスト・データベースというのは作成されておりまして、研究や教育に使われております。近年は研究用ということだけではなくて、製品としても一般向けの商品が売り出されているという状態であります。

日本では、なぜこのデータベースの開発が遅れているかということをいろいろ考えているのですが、社会科学やもちろん理科系では、コンピュータが非常に使われております。けれども、文学、あるいは国語研究というところではそもそもコンピュータがあまり使われていない、ということも大きな原因です。それともう1つやはり、著作権ということを重要視しないという風土が大きな原因になっているのではないかと思います。コピー、真似をするということにあまり日本人は罪悪感を持っていないとよくいわれますが、こういった文化的状況を別にいたしましても、あまりにも制作者の努力を評価しないことが原因になっていると思います。今でも、電子媒体に乗ったものはすべてパブリックドメインにした

†城西国際大学

ほうがいいというぐらいの暴論もあるくらいで、あまり著作権を尊重するという風土がないということは皆さんご承知のとおりだと思います。

一般に、近頃は日本研究が盛んになっておりまして、海外で各国の文学がテキスト・データベース化されていることから、日本にもテキスト・データベースがいっぱいあるだろう、ハイテク王国の日本なんだからたくさん日本の古典はデータベースになっているんだろうということで、ぜひ使いたいという要望がよく来るんですが、残念ながら対応出来るものがあんまりございません。きょうは、製作にまわっていらっしゃる貴重な方をゲストに迎えておりますが、はっきりいって、日本古典のデータベースは少ないということです。

海外の動向について早速お話ししたいと思いますが、さっきも申しましたが、1960年代から始まっていることで、今はすでに、コーパスと呼ばれる大規模なテキスト・データベースの作成が盛んに行なわれております。

サービスに関してましても非常に多様化しておりまして、フロッピーディスク、あるいはCD-ROM、磁気テープ、または電子メールなどでもサービスが行なわれているという状況です。資料に、言語別に公開を行っている研究所の名前を挙げましたけれども、そこに挙げましたのはごく一部でございまして、それらの研究所同士でネットワークもつながっております。ですから、アメリカのデータベースをイギリスに住んでいて取り寄せるというようなことも、イギリスのインスティテューションを通して取り寄せるということも可能になっております。

私の『源氏物語』は、オックスフォード・ユニバーシティ・コンピューティング・サービスに委託したわけですが、ここの場合は、実際にテキストデータベースを作成すると同時に、収集し、私のように、個人が作ったものを引き受けてサービスをするというようなこともやっております。言語別にいろいろなデータベースを集めておりまして、すでに 900 冊以上がアーカイフになっておりまして、たまたま『源氏』は日本語テキストとして、海外に公開される第1号ということになっております。

このように盛んに使われているんですけれども、一般の方々は、どうしてこんなものが必要なのかと、テキスト・データベースを作っていったい何に使うのかと疑問に思われる方のほうが多いのではないでしょうか。実際にはいろいろ使われておりまして、海外で使われている例を少しレジメに紹介いたしました。

1番目の「保存」というのは、一時、酸性紙で紙がだめになるという噂も出ましたが、それ以外にも、いずれにせよ保存するのに場所を食わないというようなことで盛んに作られております。

それからまた、欧米は、「知識は力なり」といった伝統的な考え方がありまして、文芸復興のルネッサンスを例に挙げるまでもないのですが、古典の中に人間の英知が蓄積されており、こういうものはきちんと守るべきだという発想があって、まず、保存ということが始まったんだと思います。

それから、フランスのように、自分の国の言語を守るということから、17世紀、18世紀の小説を全部入力してしまうというようなこともあります。オックスフォードの場合は、さきほども申しましたが、収集もやっております。大英帝国の名残ではありませんけれども、世界中の英知を集めるということに非常に熱心です。

2番目の、「研究」ですけれども、研究でいちばん華々しいのは、皆さんも新聞などでお読みになったかと思いますが、シェークスピアの研究であります。シェークスピアがほんとに書いたかということがいろいろ問題視されて、ベーコンが書いたとか、女房のアン・ハザウェーが書いたとかいろいろな説がありまして、そういうものをコンピュータで研究するということで成果が上がってくるというわけです。

それから、「語彙研究」などについてですが、たとえば、フランスのいろいろな詩人の、ベルレーヌとかランボーの詩の全部を入れて語彙を研究するというようなことに盛んに使われております。

つぎに、「写本研究」といいますのは、特に古いものになりますと、沢山の写本が出来ているわけです。ですから、いろいろなバージョンがあるということで、そのバージョンの違いによってそれぞれの研究の基礎が変わってくることになるわけです。それらのバージョンの研究にコンピュータが役に立つわけです。チョーサーなどは、55 個ぐらいの写本のバージョンを全部入力して、その間を飛び交いして、どのバージョンではどのように書かれているかということがすぐ見られるようなハイパーテキストも出来ております。

その他、古英語の詩で有名な『ベーウルフ』とい

うのがあるのですが、これもまた、古文書の字もグラフィックスに入っておりますし、それの翻訳、それから現代英語に読み替えた校訂の本も全部入っております。

それから、聖書に関しては大体3メガぐらいなんですが、フロッピーでサービスされておりまして、英訳だけで130ポンドぐらいで売られていおります。全部ヘブル語、ギリシャ語全体を入れると200ポンドというような形で、研究者、それから一般の人にも売られております。

また、聖書に関しては特別かと思いますが、オンラインでもサービスがされております。オンラインは30ポンドと安いんですけれども、これは、Eメールでサービスをするということで、キャラクター文字の制限があるということから安いのかと思います。これなんかは、コピーライトの所有者が面白い条件をつけておりまして、教会に1ヵ所必ず寄付してくれというような条件をつけて、公開をしているというようなこともあります。

それから、英語の詩については、ケンブリッジにある電子化出版の会社で大きい所が全部入れておりまして、これはCD-ROMになっていると思います。英国では散文よりは詩のほうが文学的にはレベルが高いという風潮がありますので、非常に高価ですけれども、2,400ポンドで売り出されております。そういった具合に、公開すると言っても、きちんとおカネを取って研究者にサービスしているわけで、そういった状況であります。

それから、教育ということについては、たとえば、ディケンズやシェークスピアのテキストなど、それぞれ好きなテキストを持って、コンピュータルームに行って文章解析の授業をやるというようなことも行なわれております。日本では、コンピュータルームで『源氏物語』の授業をやったり『枕草子』の授業をやるというようなことは、今のところ夢のまた夢という感じですが、文章解析の授業がカリキュラムに組まれているというような状況です。

それから、「電子化辞書の開発」にどういう影響があるかといいますと、辞書には、いろいろな語の出典について、「引用」をつけなければなりません。そういうことで、たくさんテキスト・データベースがあったほうが辞書が豊富になるわけで、こういった辞書開発にもずいぶん利用されております。

電子化辞書の代表的なものは、もちろん皆さんはご存じだと思いますが、『オックスフォード・イングリッシュ・ディクショナリィ』です。20巻ある大きなものが、CD-ROMになって大体500ポンドで売られております。そのほか英語の辞書の場合は、コリンズ、あるいはロングマンと幾つもの辞書会社が電子化辞書を発売しております。

それから、「ハイパーテキスト」で、いちばん有名なのはハーバード大学が開発した『ペルセウス』というテキストです。これはギリシャの文化、テキスト、それに基づくコメンタリー、翻訳、それからグラフィックスとすべてが入っております。これは10年のプロジェクトだったと思いますが、こういったものが作られております。これはサイトライセンスだけだと思いますけれども、100万円ぐらいで各研究所に公開されております。

それから、これはたまたま先週のことなんですが、イタリアからトマス・アキナスのCD-ROMが出来たから使わないかといってきました。これは30年のプロジェクトだそうです。公開と言ってもやはり売っているわけで、994ドル、約千ドルの値段で公開がされております。

れから、私がプラトンのテキストを買いましたカリフォルニア大学の付属機関、TLGと簡単にいっておりますが、TLGは、個人のユーザーにはCD-ROMで、やはりこれも約500ドルで公開しています。サイトライセンスになりますと850ドルということです。ギリシャやラテンの研究をする人がそんなにいるかとお思いかもしれませんが、1年に2回ずつニュースレターが来まして、どれだけユーザーが使っているかという報告もよくしてくれるんです。それによりますと、米国ではすでに583件も使われております。イタリアで130件、ドイツでは113件、フランス、英国もそれぞれ50件ずつぐらい使っているということで、結構高い費用はかかりますが、研究所で購入をしてそれを使って研究をしているというような状況であります。

コピーライトの処理なんですけれども、1960年代からすでに入力が始まっていると申しましたが、学会活動もそのころから始まっておりまして、こういった電子化テキストを作る場合は、1人で作るのは大変ですからたいていプロジェクトになっておりまして、コピーライトの交渉も学会があたって、ネゴシエーションを本屋とやるというようなことも行なわれています。

それから、ファンドですが、カーネギーとか、あるいは国のファンド、個人のファンド、ドネーションで運営されるということが多く、国家的なプロジェクトであることも多いといわれております。

サンプルとして、TLGとそれからオックスフォードのコピーライト契約をそこに持ってまいりました。さきほど名和先生がおっしゃられたと思いますけれども、私は著作権のプロではありませんのであまりくわしくはございませんが、実際には、どこかに著作権法というのがきちんとあって、その何条かに基づいてこういうのが作られているんだと思いますけれども、実際にテキストにこういうものがついてきたので、そこのところだけをお話しするわけです。資料の2のTLGの場合は、研究用と教育用の両方に使え、もちろん非営利目的で研究と教育に使えます。

それから、サードパーティですね、第三者にコピーをしたり、あるいはトランスファすることは、あらかじめＴＬＧの許可なくしてはやってはいけないことになっております。研究者がそれを使って研究をする、そしてそれを出版するということは妨げないということですから、それを使って、たとえば二次データベース、インデックスとかコンコーダンスを作ることはもちろん構わないということです。

大事なことは、最後のところにあります。たとえテキスト・データベースを補助的に使っても、あるいは、基礎の分析のために一部を使っただけでも、それを補助として使った場合も基礎的に使った場合でも、どちらも必ずアクノレジメントをつけることです。実際にＴＬＧのこのテキスト・データベースを使ったということを明記するということが義務づけられております。

オックスフォードの場合も、これはユーザーズ・デクラレーションですか、申告という形で利用を申請するわけで、一種の規約みたいなもので、この場合は、研究利用のみです。教育利用の場合は、金額がもっと高くなって別になります。

こちらも同様にデッドコピーですが、全体をコピーした時にはもちろん、あるいは一部を使った場合でも、必ずアクノレジメントをオックスフォードのアーカイフと同時に、それからオリジナル・ディポジター、すなわち、さっきも申しましたけれども、実際にオックスフォードが作っていないテキストを委託して、サービスする場合は、実際作った人にもアクノレジメントを付けるようにと、こういう2つの条件がついております。もちろんコピーライドはオックスフォード側にあるわけで、第三者にはコピーしてはいけないことになっています。

またそのアクセスも、実際には個人だけなのですが、常識的には協同研究者の2、3人ぐらいまではオーケーだというようなことです。

それから、必ず手で入力したり、OCRを使うにしてもエラーがあるわけで、テキスト・データベースの精度ということは常に問題になりまして、必ず誤りがあったら申請するようにというようなことが明記されております。

値段に関しましては、これはオックスホードの場合はTLGよりはずっと安くて、大体、マグネティックテープでサービスされますから、英国以外だと30ポンドぐらいでサービスしてくれます。こういった具合に、ユーザー契約といいますか、コピーライトのレギュレーションをくっつけてサービスをしているというのが現状です。

海外の場合は、すでに昔から辞書学とか文献学とか書誌学ということが盛んに行なわれていて、テキストを研究するという学問が面々と続いているわけで、そういう現場にいる方たちが、コンピュータを使って今はいろいろな優れたデータベースを作るための研究を行っております。データベースをどのように作ったら電子化辞書や、あるいはハイパーテキストにも使われるような汎用性の高いものが出来るのかといったようなことが研究されておりますし、また、カンファレンスも非常に多く、勉強会のようなものが非常に多いということが、これまたデータベース作成を熱心にさせている原因かと思います。

ちょっと時間がなくなってきましたので、最後に、簡単に問題点のところをいっておきます。今回、日本のテキスト・データベースがあまり発展しないということから始まっているわけです。本がない場合については、会場にいらっしゃるご専門家の方にあとでお話しいただけるのではないかと思いますが、マニュースクリプト、あるいは古文書を本にしないで、そのままコンピュータに入れる場合です。

一般には、本がある場合は、それをもとにテキスト・データベースを作るということが行なわれているわけです。しかし、これまでは大変な作業でありました。今もトマス・アキナスの話をしましたが、30年プロジェクトなどというのがあるわけで、大

変であります。ギリシャ文字の場合は、ローマ文字と数字に読み換えるというような作業をして入力するとか、いろいろな、入力の際に実際の本とはずいぶん違った形で工夫をしなければいけない。そういうことから、実際には製作者の側にテキスト・データベースの権利があるということがいわれています。そして、実際に出版者側がどんな権利を持っているか、そして、作った側がどんな権利を持っているかについては、これから会場の方から教えていただこうと思っておりますけれども、そういうところが日本では問題になってくるのではないかと思っています。

海外の場合は、出版社は、電子化されたテキストがあったほうが本が売れると考えるということで、割に、「電子化テキストを作りたい」と言うと、ジェネラスに許可をくれるというようなことがあるということです。また、出版社自身が、本を出すのと同時に、テキスト・データベースのほうも作る。そして、研究が盛んになれば、また、もっと本も売れるというようなことを考えているのだそうで、そういう意味では、ますますコマーシャルなほうも盛んに作られております。

サービスに関しては、日本ではどういったことが障害になっているかというようなことはあとでご議論願いたいと思います。「将来」ということを書きましたのは、これまでは、今申し上げましたように何十年プロジェクトといわれるように非常に苦労して作るわけですが、ＯＣＲがほとんど発達してきますと、今のゼロックスのコピーのように簡単にコピー出来るようになってまいります。そうなってくると、テキスト・データベースにどういった付加価値をつけていくと優れたデータベースになるのかというようなことも問題になってくるわけです。

さっきのチャドウィックヒーリー社の所が出しているものに、もう１つ「ラテン教父のテキスト」というのがあるんですけれども、これも非常に高いものなんですが、もうすでにSGMLで入力されているというようなことが書いてありまして、それが宣伝になっています。ですから、付加価値をどのようにつけるかということでそれだけテキストの値段も上がっているというような状況です。ですから、こういうSGMLのような付加価値も、著作権にするかという問題も出てくるかと思います。

それから、今後は文字だけでなくてグラフィックスも入れたような、さっきも申しました中世のテキストの場合、古文書の実際に出てくる文字を写真で撮ってそれをグラフィックスに読み込んで、それも見れるようになっている。そうなってくると、グラフィックスなんかとほかのメディアとの重複といったようなことが起きてくるとします。そうなると著作権の問題がどうなるのか。将来のことについていろいろと疑問があるわけで、ですから、問題点は私は前座として提起するのが役目ですので一応挙げておきましたが、これ以外にもたくさんあるかと思います。こういうところにポイントがあるのではないかということで終わらせていただきます。

どうもありがとうございました。（拍手）

司会：ありがとうございました。ご質問もあろうかと思いますが、とりあえずきょう、発表者のご意見を全部伺ってしまうというようにしたいと思います。では、２番目の安永先生お願いいたします。どうぞ。