特　集　「研究開発における情報利用と著作権」

コーパス構築と著作権保護

Copyright protection and corpus development

前川 喜久雄　国立国語研究所言語資源研究系
Kikuo Maekawa　Dept. Corpus Studies. National Institute for Japanese Language and Linguistics
kikuo@ninjal.ac.jp

Keywords: language corpus, balanced corpus, speech and language analysis, copyright

1. はじめに

筆者の専門は音声学であるが、10 年ほど前に音声自動認識のための大規模な話し言葉コーパスである『日本語話し言葉コーパス』の構築に携わる機会があり[前川 04]、それを契機としてコーパスの構築が仕事のなかで大きな比重を占めるようになった。2006 年からは科研費特定領域研究「日本語コーパス」[前川 09]の領域代表者を務めている。このプロジェクトでは日本語に関する初の均衡コーパスである『現代日本語書き言葉均衡コーパス』の構築に関わっており、わけても著作権処理作業に深く関係している。

本稿では、我が国の著作権の問題点をコーパス構築の立場から経験的に指摘する。編集部から与えられたタイトルは「言語解析と著作権」だったが、ご覧のとおり少し修正させていただいた。

2. BCCWJ の著作権処理

現代日本語のコーパス言語学的研究を行う際の大きな問題は均衡コーパスが存在しないことである。この問題を解決するために『現代日本語書き言葉均衡コーパス』(Balanced Corpus of Contemporary Written Japanese, 以下 BCCWJ と略する）の構築を始めた。均衡コーパス(balanced corpus)とは、対象言語の全体がバランスよく反映され、その言語の正確な縮図となっているコーパスを意味する。しかし各国語の均衡コーパスとされているものの設計を検討しても、均衡コーパスを設計する方法が確立されているわけではない。

BCCWJ では、日本語の全体像を可能な限り正確に把握するための方策として、できる限り無作為抽出によってサンプルを選択することにした。書籍、新聞、雑誌など母集団となりうる出版データが公開されている場合は厳密な無作為抽出を実施している。出版に関する母集団だけでなく、東京都の公立図書館に所蔵されている書籍全体を母集団とした無作為抽出もおこなっている。

BCCWJ の規模は全体で 1 億語であるが、そのうち約 6000 万語が書籍のデータであり、約 24000 件の著作権処理が必要になった。BCCWJ の構築を一般企業が行っていたとしたら、この段階で著作権処理を諦めるだろう。

実は我々も著作権処理を回避する可能性がないかについて専門家に意見をもとめた。著作権法上の「引用」として処理できるかどうかが議論の焦点となったが、結局、国の機関がやることだから、きちんと優等生的な処理をしようということになった。5 年という短期間で膨大な著作権処理ができるものか、成算などありはしないが、プロジェクトの成否がここにかかっているのは確かだから、ともかく始めることにした。

最初に取り組んだのが、日本文藝家協会をはじめとする著作権管理団体との交渉である。幸い各団体の協力を得ることができ、合計で約 4000 名に協力依頼状を送ることができた。もし貴殿の作品がコーパスの対象となった場合、無償でサンプルとさせていただきたいという依頼である。会員からの反応も良好であったが、実際に無作為抽出されたサンプルとつきあわせてみると、この方法で著作権を処理できるのは書籍サンプルの 3%程度であることがわかった。残るサンプルについては、地道に努力するしかない。職員 4 名とアルバイトからなるチームを作って処理を開始した。

この著作権処理で最も大変なのは、権利者の連絡先を見つける作業である。著作権には登録制度がないからインターネットや各種有償データベース、各界の紳士録などを駆使して権利者(とおぼしい人)の連絡先を調べる。

書籍なのだから、出版社に連絡をとればよいと思うかもしれない。しかし、そう簡単にはことが進まない。

まず 2005 年に個人情報保護法が完全実施されたことによって、大手出版社は勝手に著作者の情報を外部に提供することにできなくなった。会社としては「国語研からこのような依頼が来ていますが、貴殿の連絡先を教えてよろしいでしょうか」というお伺いをたてて、許可をもらった後でなければ、連絡先を教えられないのである。

それに大手出版社の場合、数百件のサンプルがあたる

ことが珍しくないから、そんな面倒につきあってくれる会社はあまりない。あまりないということは、その面倒につきあってくれた出版社があったことを含意している。社名を挙げることはしないが涙がでるほど嬉しかった。ただしこういう奇特な会社はやはり少数派である。

連絡先が判明した権利者には、実際にどの著作のどこからどこまでを利用したいかの情報とともに著作物無償利用の依頼状を発送する。図1に書籍からとられたサンプルに関する著作権処理の現状を示した。サンプルの総数は24050件であり、上の円グラフは連絡先の調査状況を示している。現在までに8割強のサンプルの連絡先が判明しており、最終的には多分9割のサンプルの連絡先が判明するだろうと予想している。「連絡不能」となっているのは、一定期間調査しても連絡先が見つからなかったり、発送した依頼状が配送不能で返ってきたりした場合である。

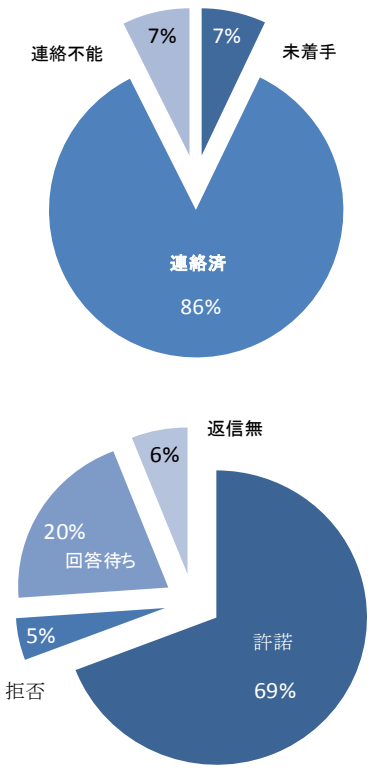

図1 書籍サンプルの著作権処理の現状.

下のグラフは上のグラフで「連絡済」となっているサンプル(20574件)に対する権利者の反応の分類である。許諾がもらえたのは約7割の権利者からである、では残りは拒否かというとそうではない。明示的な拒否は5%にすぎない。最も多い反応は、いつまでたっても回答がもらえない場合である。図では発送後一定期間内のサンプルは「回答待ち」に分類しており、一定期間を過ぎて催促しても反応のないサンプルは「返信無」に分類している。

雑誌サンプル（2350件）の状況は少し異なる。連絡済が57%、そのうち許諾が61%、拒否が7%であり、書籍よりも許諾率が悪い。その理由は4節で推測する。

3. 法改正との関係

3.1. 連絡不能・未回答サンプルの問題

BCCWJの著作権処理には本稿の執筆時点で半年ほどの時間が残されている。図1に示した状況が今後も続くとすれば、サンプル全体の9割程度が連絡済となり、その7割程度に許諾が得られるだろう。結局サンプル全体に対する許諾率は6割強にとどまるので、残る3割強のサンプルをどう扱うかが問題である。明示的に拒否されたサンプルは当然公開対象からはずすことになるので悩む必要がない。悩ましいのは最後まで連絡済とすることができなかったサンプルと回答が得られなかったサンプルである。この種のサンプルは全体の3割弱を占めるはずである。

ここで著作権処理の王道は、文化庁長官の裁定をあおぐことである。しかし従来の長官裁定制度は著作権利用者の側からみるといろいろと不透明な部分がある制度で、あまり使う気になれなかった（実際これまでは平均して年1回ぐらいしか利用されてきていないはずである）。

大手出版社が過去の書籍の再刊やアンソロジーの出版などで権利者に連絡がとれない場合は、その旨を本の扉などに印刷した上で(あるいはそういうこともせずに)出版してしまうことが多いようである。

今回の改正では、裁定が下るまでの期間に著作物を仮利用してよいことになっている。これはたしかに利用者の側に立った改正であり、大変ありがたい。しかし、まだ問題が残っている。裁定制度を利用するには供託金を収めるのだが、その金額次第では制度を利用したくとも利用できなくなるのだが、供託金は申請の後で検討される仕組みになっている。

これはおそらく利用の状況（出版部数など）に応じて金額を決めようという良心的な配慮に基づく制度なのだろうが、BCCWJのように裁定をうけるべきサンプルが何千件にも達する場合、申請しないと金額がわからないのでは困る。大雑把な目安で分野ごとに一律の金額を決められないものだろうか。

3.2. 検索エンジンに関わる権利制限

BCCWJにはインターネット掲示板である Yahoo!知恵袋と Yahoo!ブログのデータが1000万語ずつ格納されており、コーパスの価値を高めている。これらのデータを公開できたのは、ヤフー株式会社が著作権問題の解決に積極的に協力してくださったからであり、筆者らはこれらのデータの著作権処理を主体的には行っていない。

ところで、今回の法改正にはインターネット検索エンジンに関わる権利制限の導入が含まれており（権利制限というのは、これまで保護されてきた著作者側の権利を一部制限するという意味であり、検索エンジンの利用を制限するのではないことに注意）、改正の目玉となっている。この改正によって、ネット上のテキスト等を一時的にサーバーに保存した状態で解析し、その成果を公表することが可能になった。これによってクローリングデータの研究利用がかなり広い範囲で安心して行えるようになることは確かである。

しかし、解析のために収集したデータを公開してよいかどうかについては言及がないので、問題はグレーゾーンに残されたままである。実は法改正に先だつパブリックコメントの段階までは、改正案に学術目的でのデータ公開を可能にしたと解釈できる条文が含まれていた。それがどのような経緯で消えることになったのか詳らかにしないが、研究用データ公開の必要性は、2008 年夏に筆者が文化審議会著作権分科会法制小委員会で参考人として意見を述べた際に、クローリングデータの解析目的での収集保存とならんで強調した点であっただけに、残念な後退である。

3.3. フェアユース

今回の法改正にいたる文化審議会での審議の中では、いわゆる日本版フェアユース条項の導入が真剣に検討された。一時、導入が決まったかの報道がなされたこともあったが、実際は導入に到らなかった。

筆者なりに説明すると、フェアユースとは、現在の著作権法がそうであるように権利制限のケースをひとつずつ具体的に列挙するのではなく、抽象的な原則だけを示すことにして、個々のケースが権利制限の対象にあたるかどうかの判断は利用者に委ねる制度である。利用者は法律に示された原則に照らして公正な利用であると判断すれば、権利者の許諾を得ることなく著作物を利用すればよい。それを公正な利用と認めたくない権利者は、訴訟を起こして、裁判で黒白をはっきりさせることになる。

先に示した BCCWJ の著作権処理の現状からは、訴訟を起こす可能性がある権利者は最大で 5%前後と推測される。これを少ないとみるか多いと見るかは意見の分かれるところだろう。前述の法制小委員会でも、筆者の説明に対し、5%は少ないとは言えないと発言した委員がおられた。

しかし筆者はフェアユース条項の導入は長期的には必然であろうと考えている。現在の著作権法は、著作物の権利者が社会全体からみれば少数のプロフェッショナルに限られている時代に作られたものであるため、インターネットに代表される安価な情報発信手段の爆発的普及によって国民の多くが著作権者と化した情報化社会（それは仮想社会の話ではなく、今私たちの目の前に広がっている現実である）においては、もはや個別的な制限規定では急速に変化（あえて進化とは言わない）しつづける現実に追いつくことができないからである。

BCCWJ の著作権処理では、現行法の定めに従ってあえて愚直に処理を進めたが、本当はコーパスのための著作権処理そのものが、現行法にとって想定外の事態と言うべきであろう。

今になってみると、筆者らの著作権処理作業には著作権に関する一種の社会実験としての意義があった。この作業を通じてふたつの重要問題についての知見が得られたからである。ひとつは無作為に抽出された著作物の権利者にどの程度接触することができるかという問題。これにはどう頑張っても 8 割までという知見が得られた。

もうひとつは、学術目的での著作物の無償利用に対する国民の反応がどのようなものかという問題。これには、筆者らの依頼に対して何らかの意思を表明した権利者（図 1 下側のグラフから回答待ちと返信無しを除外した残り）のうち、85%が明示的に利用を許諾し、明示的な拒否は 6%強という知見が得られた。国民の大半は学術目的での無償利用について十分に寛容である。この結果をみたうえで、あえて 6%を重視するのは、為政者にとって賢明なふるまいとは言えないだろう。

4. 拒否の理由

BCCWJ の著作権処理の現状を紹介すると、よく受ける質問がある。拒否の理由である。これは我々も気になるところなので、できるだけ情報を収集するようにしている。理由の類型化は難しいが、書籍サンプルについて主要な理由を列挙してみる。

権利者のなかにはテキストを修正したいという要望を伝えてこられる方が少なくない。コーパスでの利用は構わないのだが、文章に手をいれたいという希望である。単純な誤植であれば、BCCWJ には誤植のありかを示して正しい文字列を示すタグが用意されているから、それで対応できるのだが、誤植以外の修正となると認めるわけにはいかない。コーパスの主旨を説明しても納得していただけない場合は利用を諦めざるをえなくなる。同じ作品の別の箇所ならばと先方から提案されることもあるが、これを認めるときりがなくなるので、涙をのんでお断りしている。類例として、昔出版したことはしたが、今となっては自分で納得がゆかない文章だからコーパスを介してでも人の目に触れるようにしたくないという方もおられた。

作品の一部だけを利用されるのは嫌だ。作品全体ならば許諾してよいという方もおられた。可変長サンプルの存在を説明するが、それで納得していただけない場合は諦めることになる。

以上は、いわば表現者としての矜持による拒否であるから、断られても仕方がないという気がする。しかし当

然のことながら、そういうケースばかりではない。著作権処理も交渉ごとであるから、権利者と処理担当者との相性のようなものが存在する。担当者の電話での口の聞き方が気に入らない、依頼文書が何を言いたいのかわけがわからない、研究所代表に電話をかけたらたらいまわしにされた、といった理由で拒否されることがある。

明らかな誤解や思いこみもある。コーパスを模範的日本語の集積と考えて、国の機関が文章の良し悪しを決めるなどおこがましいとい叱られたことがある。逆に無作為抽出とはなにごとかと叱られたこともある。世の中の日本語は間違いだらけなのに、それを集めて国が公開するなどもってのほかという論旨である。そういえば、国の機関が国民の著作物を利用するのに無償利用するのは憲法違反だといわれたこともあった。

東大文学部の元教授と名乗る方から、弟子の文章を私に無断で利用するとは何事か、謝罪せよというファックスが突然送りつけられてきたこともあったが、これなど一体何をどのように誤解しているのか理解に苦しんだ。

以上は書籍サンプルに関する個人の権利者の反応である。一方、法人が著作権を有しているサンプルもある。その典型は雑誌のサンプルである。雑誌サンプルの許諾率が書籍よりも低いことは先に述べたが、個人としての判断に比べると、組織の一員としての判断はどうしても安全サイドに偏るのは避けがたいようである。著作権が法務部等で一括管理されておらず、許諾の決定を個々の編集部に委ねてしまう会社が多いことも拒否回答を増加させる要因であると思われる。

出版社の判断に触れたところで、是非論じておきたい問題がある。「版権」の問題である。書籍には版権、すなわちその本を出版する権利があり、それは出版社が保有している。従って著作権処理では著作権者からの許可に加えて、版権を有する出版社からも許可をとらなければならない。そう主張する会社がたまにある。

しかし日本の著作権法には「版権」という言葉は書かれていない。「出版権」ならば第79~83条に規定されているが、これはある書籍の全体をそのまま印刷出版する権利であり、コーパスのように作品の一部だけを利用するケースは該当しない（作品全体であっても電子出版ならば該当しないという解釈もある[作花 04]p.432参照）。さらに80条3項は「出版権者は他人に対し、その出版権の目的である著作物の複製を許諾することができない」と述べて、サブライセンスを禁止している。

以上のように法律上の出版権は、筆者らの行っている著作権処理に対して、出版社が利用を拒否する根拠にはならない。そして、いわゆる「版権」は出版界の慣習であって法律上の権利ではない。しかし、一部とはいいながら「版権」が出版社の既得権であると強く主張する出版関係者がおられるのは残念なことである。

本節には否定的な意見ばかりを紹介することになったが、これらが少数意見であることを最後に強調しておきたい。少なくとも個人権利者に関しては、コーパスの構築に賛成だ、完成を楽しみにしているといった意見の方が圧倒的に多い。許諾上に添えられてくるそのような言葉が筆者らの仕事の励みになっている。

5. まとめ（国民の理解を得るために）

以上、筆者の経験に基づいてコーパス構築に伴う著作権の問題を論じた。コーパス構築の観点からすると、今回の改正を経た著作権法にも問題は多く、最終的には日本版フェアユース条項の導入が望まれる状況にある。

前節で紹介したように、著作権に対する国民の理解ないし感情には様々なものがある。それは現実であるからそのまま受け入れるしかないのだが、少しでも多くの国民の理解を得るために、我々研究者も自らの研究内容を分かりやすく説明する必要がある。

かつて著作権課に席をおいた人から聞いた話では、著作権課にくる陳情はほぼすべてが権利拡大の要求であり、権利制限に関する陳情は無いに等しかったそうである。

我々も著作権法の改正が必要と考えるならばその旨を率直に発言して関係者を説得すべきであろう。タイミングを見計らって関連学会が連盟で文化庁著作権課に要望書を提出するようなことも遠慮すべきではない。

学術は原則として行政や政治から独立しているべきであるが、昨今の社会ではその独立性を保つためにこそ、発言が必要とされるからである。

そのような状況の一典型例を付録に述べたので参照していただきたい。

付録（コーパスは誰でもできる？）

筆者の属する国立国語研究書は2009年9月末をもって文化庁所属の独立行政法人から大学共同利用機関へと移管された。この移管に係る検討の過程で筆者らの仕事の価値が一旦は政府によって全否定されるという事態が生じた。著作権の問題とは直接関係しないが、ここでその経緯を報告しておきたい。政府が主催する委員会の有識者といえどもその見識が広いとは限らないこと、従って研究者は日常的に自らの仕事の意義を国民全般に対して説明しておかないと、とんでもない誤解を受ける可能性があることを実例をもって示すためである。

現在、民主党主導による、独立行政法人とその業務の「仕分け」が進められているが、独立行政法人業務の見直しを任務とする委員会は自民党政権化でも活動していた。渡辺喜美行政改革担当大臣を担当閣僚とする行政減量・効率化有識者会議（座長はキッコーマン会長の茂木友三郎氏）である。民主党による仕分けとの相違点はインターネットによる中継がなく、政治ショーと化してい

なかった点であるが、実はこれが問題だった。民主党の仕分けは「人民裁判」と揶揄されたが、自民党による独立行政法人改変は「秘密裁判」だったからである。

後日入手した有識者会議の詳細な議事録を読むと、審議の過程で一部の委員がコーパスとは何かを理解せず、思い込みによる否定的発言を繰り返している。「データベースは誰でもできることであり，そのような事業にいたずらに国費を投入しなくてもよい。」「国語のデータベースなんて世間にはいくらでもあることで，10 人くらいでやっていることなら外に切り出せるだろうから，大学なんかでも十分できるでしょう。」というのは安念潤司氏（中央大学法科大学院教授、当事）の発言。何故データベース（コーパスのこと）が誰でもできると判断できるかの根拠は一切示されていない。

「言葉を扱う新聞社で仕事をしてきたが，これまで国立国語研究所に世話になったことは何もなく，新聞協会とかでやっているもので十分間に合っている。国立国語研究所は，理解不能なものの対応や国民に通用しない日本語を研究しているとしか思えない。これはもう国立国語研究所そのものを廃止することが一番いいことではないかと思いますね。外に行った時に役立つように海外で日本語を普及するようなことならまだしも，11 億円もかけて部屋にこもってデータベースを作ってもらっても何も有り難くない。廃止していただければ有り難いですね。いや廃止です。」は菊池哲郎氏（毎日新聞社常務取締役、当事）の発言。

この発言がなされた時期には文化審議会国語分科会漢字小委員会による常用漢字見直し作業が進んでいた。そこでは漢字の取捨選択における基礎データとして小委員会が依拠したデータに対して疑義が表明され、他ならぬ新聞協会によって、国語研が開発中の BCCWJ を利用してはどうかという提案がなされていた(2008 年には実際にデータを提供した)。菊地氏はそのことを知っていて無視したのかどうか。また 11 億円というのは国語研の運営費交付金の総額であり、コーパス構築に用いているのは 1 億円程度であるのだが、ヒアリングの席に呼ばれていた文化庁関係者はすっかり萎縮していたようで、この誤りを訂正した様子がない。

このような審議を経て有識者は 2007 年 11 月に国立国語研究所に関する勧告案の骨子をまとめるが、そこには「コーパス事業は民間でもやっているので国語研が自ら行う必要はない。廃止又は仕様を決めて入札に出すべき。」とあった。安念氏、菊地氏の意見そのままである。

筆者らが有識者会議の審議内容を知ったのは、この時が最初であり、全くの寝耳に水状態であった。とりあえず筆者は文化庁国語課を介して民間での実施とは具体的に何をさしているのかを明らかにしてほしいと有識者会議に要求した。これには「小学館コーパスネットワーク」という回答があった。

ご存知の方も多いだろうが小学館コーパスネットワークは英語の均衡コーパスとして有名な British National Corpus その他の英語コーパスを日本国内で配信するサービスである。いくらなんでも英語のコーパスがあれば日本語コーパスは不要と考えているわけはないだろうから、実際のサービス内容を確認せずに名称だけで日本語コーパスと誤解した可能性が高い。有識者会議には最初から結論があり、その結論を正当化するために日本の民間会社の「コーパス」を検索してみたという構図がすけてみえる。

この回答に対し筆者はもちろん意見書を書いた。しかしその文書は文化庁国語課から有識者会議に渡っていなかったことが後日発覚した。一説には内容を読んだ有識者会議から受けとりを拒否され、それを取引成立と解釈した文化庁の役人が文書を持ちかえったとも言われているが、そのような取引が成立していなかったことは同年 12 月 24 日の閣議決定（福田康夫内閣）で独立行政法人国立国語研究所を廃止し、大学共同利用機関に移管することが決定されたことから明らかである。

その後 2009 年 3 月に国立国語研究所を大学共同利用機関に移管するための法律改正に際し、衆参両院で従来の国語研究所の機能を保全すべしとの附帯決議がついた。また移管後の研究所のあり方を検討した科学技術・学術審議会において独法時代のコーパス事業が高い評価を受けたこともあって、現在の国立国語研究所には言語資源研究系に加えてコーパス開発センターが設置されている。

これによってコーパス構築が誰でもできる仕事ではないことが公に認められたものと解釈したい。

◇ 参考文献 ◇

[作花 04] 作花文雄：詳解著作権法（第 3 版），ぎょうせい(2004).
[前川 04] 前川喜久雄：『日本語話し言葉コーパス』の概要，日本語科学, No.15, pp.111-133 (2004).
[前川 09]前川喜久雄：代表性を有する大規模日本語書き言葉コーパスの構築，人工知能学会誌, Vol.24, No.5, pp.616-622 (2009).

2010 年 **月 **日　受理

著 者 紹 介

前川　喜久雄（非会員）
1956 年生. 1980 年上智大学大学院博士後期課程（言語学）退学. 専門は音声学であるが，自発音声研究のために『日本語話し言葉コーパス』の構築に関係したことから言語資源学が第二の専門となった．大学共同利用機関人間文化研究機構国立国語研究所教授，言語資源研究系長.