コーパス検索と著作権

田淵 龍二（ミント音声教育研究所）

キーワード： コーパス，検索エンジン，著作権，言語資源，公開

1. はじめに

コーパスを使った研究や学習の機会が増大しているが，利用者にとっては所望の情報に円滑にアクセスするための検索システムが必須で，たとえば Google や Yahoo などの検索サイトが有名である。そうした検索エンジンの開発が盛んで，科研費研究の対象分野でもある。他方，コーパスは著作物であることから著作権の問題が生起するが，IT 時代に法律が対応しきれておらず混乱や萎縮が心配され，「教育機関の著作権法に関する理解が不十分」（文化庁長官官房著作権課, 2017）との意見もある。

ところで，現在，国において柔軟な権利制限規定を盛り込む著作権法改正が検討されている。その資料では，所在検索サービス（広く公衆がアクセス可能な情報の所在を検索可能にするとともに，その一部を検索結果と併せて表示するサービス）にあっては「インターネットにアップロードされている著作物に限れば，現在も提供可能」（文化庁長官官房著作権課, 2018）と明記されている。

2. 目的

コーパス構築とコーパス検索エンジン開発における著作権の処理実態をあきらかにする。

3. 方法

コーパス構築とコーパス検索エンジン開発の研究を行ってきた専門家が集まる言語処理学会の学会誌「自然言語処理」及び年次大会発表論文集に掲載されている論文で著作権の処理実態を調査した。対象は，1994 年の第 1 巻から第 24 巻（2017 年）までの学会誌 565 本と，2004 年から 2017 年までの発表論文集 3,863 本，あわせて 4,428 本である。文中に「コーパス」と「著作権」を含む論文を，NaCSE で検索した。NaCSE（2018）は言語処理学会の論文を検索する専用のオープンサイトである。

4. 結果

コーパスと著作権を同時に含む論文は 66 本あり，本文に共起した論文は 40 本であった。本文共起とは，同一文脈にコーパスと著作権が出現したことを意味する。これらを調査したところ，構築したコーパスにおける著作権の扱いは 7 つに分類され（表 1），扱いを決めた要因は 7 つに整理された（表 2）。

表 1.
コーパス構築における著作権の扱い

1. 著作権処理を実施して公開
2. 保護期間終了の著作物のみ公開
3. 公開を断念
4. 解析結果（例：n-gram）のみ公開
5. 元著作物へのリンクを公開
6. （疑似）テキストを生成して公開
7. クリエイティブ・コモンズで公開

表 2.
著作権の扱いを左右した主な要因

A. 著作物が自前か否か
B. 著作権保護期間内か否か
C. 著作権者が少数か膨大か
D. 著作物がウェブ公開か否か
E. 公開が目的か否か
F. 必要なのは著作物本体か解析結果か
G. 製作者の予算と人員の大小

例えば国立国語研究所は書籍だけで24,000件の著作権処理を職員4名とアルバイトのチームを編成して5年の歳月をかけて行なった（前川，2010）。同じ国立国語研究所の浅原・岡（2017）は，「国語研日本語ウェブコーパス」の公開にあたり元データへのリンクを貼ることで解決している。書籍にあってはウェブのようにリンクは貼れないので，藤田ら（2017）の絵本検索システム「ぴたりえ」では，絵本の表紙を表示するだけにとどめている。さらに山田ら（2017）は二言語並行コーパス構築にあたり日本記者クラブと覚書を締結している。中條ら（2016）はウェブコーパスで収集した英文を元に英文を自作しクリエイティブ・コモンズとして公開している。また，渋木ら（2014）のようにコーパス公開を断念したり，解析結果のみを公開する例や擬似文を自動生成する例（伊藤, 2014）も見られた。

5. 考察

コーパスと検索エンジンを公開するか否かは開発目的により異なり，公開可能か否かは個々の著作物により異なることから，一律に論じることはできない。しかし，著作権問題に萎縮して放棄あるいは禁止したり，さらには法律論に深入りするよりも，表2で見た事情に応じて，表1の対処方法を目安にすることが現実的であり，著作権法の精神「この法律は，（中略）著作者等の権利の保護を図り，もつて文化の発展に寄与することを目的とする」（第一条）に沿うものではないかと思量される。

2万件を越える著作権処理を行った前川（2010）はフェアユース導入を推奨している。フェアユースは，米国で実施されており，公正な利用であれば許諾を得ることなく利用できるとしたものである。

フェアユース同様に，許諾なく著作物を利用できる条件を著作権者自らが示したものにクリエイティブ・コモンズ・ライセンス（CCL）がある。CCLは著作物の流通促進を目的としている。例えばWikipedia, YouTube，TED，初音ミク，学術誌Natureなどが CCLを採用している。

内閣はIT時代に対応した「柔軟な権利制限規定」を含む著作権法改正案を国会に提出，可決された（5月18日）。教育メディア学術団体にあっても，世情に合わせた著作権への理解と対応が求められている。

参考文献・参考資料

浅原正幸・岡照晃 (2017).「nwjc2vec:『国語研日本語ウェブコーパス』に基づく単語の分散表現データ」言語処理学会第23回年次大会発表論文集, 94−97.

文化庁長官官房著作権課 (2017).「遠隔授業に関する著作権制度について」. Retrieved from http://www8.cao.go.jp/kisei-kaikaku/suishin/meeting/wg/toushi/20170224/170224toushi06.pdf

文化庁長官官房著作権課 (2018).「著作権法の一部を改正する法律案 概要説明資料（ＡＩの利活用促進関係）」. Retrieved from https://www.kantei.go.jp/jp/singi/titeki2/tyousakai/kensho_hyoka_kikaku/2018/sangyou/dai5/siryou2-4.pdf

中條清美・内山将夫・赤瀬川史朗・西垣知佳子 (2016).「データ駆動型英語学習における教育用例文コーパス SCoREの活用」言語処理学会第22回年次大会発表論文集, 1081−1084.

藤田早苗・服部正嗣・小林哲生・奥村優子・青山一生学 (2017).「絵本検索システム「ぴたりえ」～子どもにぴったりの絵本を見つけます～」自然言語処理, *24*, 49−93.

伊藤薫 (2014).「比喩表現コーパスの構築と問題点 -言語学の立場から-」言語処理学会第20回年次大会発表論文集, 149−152.

前川喜久雄 (2010).「コーパス構築と著作権保護」人工知能学会誌, *25*, 628−632.

NaCSE (2018). http://www.mintap.com/nacse/nacse.html

渋木英潔・中野正寛・宮崎林太郎・石下円香・金子浩一・永井隆広・森 辰則 (2014).「情報信憑性判断支援のための Web 文書向け要約生成タスクにおけるアノテーション」自然言語処理, *21*, 158−212.

山田優・松下佳世・石塚浩之・歳岡冴香・Carl Michael (2017).「記者会見通訳の二言語並行コーパスの構築」言語処理学会第23回年次大会発表論文集, 1168−1171.