論文閲覧を支援する試み ― 文脈検索可能な NLP 予稿集コーパス構築

田淵 龍二
ミント音声教育研究所
tabuchiryuji@nifty.ne.jp

1. はじめに

学術研究には，独創性や新規性が求められる。しかし何が独創的で，どこが新規なのかを知ることは簡単ではない。なぜなら，文物の創作にあたって，すべての関連分野で正統的教育を経ているとは限らない研究者が，先人の道を俯瞰することは容易ではないからだ。そこで，初級者でも先行研究から手軽に学べる道具が必要だと考えた。本研究の目的は，言語処理学会年次大会発表論文集（NLP 予稿集）の文や段落の拾い読みができる文脈検索サイトのコーパス構築である。

2. 先行研究

先行研究としては，実例としてグーグルスカラー（Google Scholar; https:// scholar.google.co.jp/ ）と J-STAGE（https://www.jstage.jst.go.jp/ ）を取り上げる。検索の動機を「コーパスにおける著作権問題」と仮定し，検索語は「コーパス」と「著作権」とした。

グーグルスカラーではヒット文献が上から下に 1 列に並ぶ。各文献には 4 つの項目があり，上から順に，題名，著者，検索語を含むテキストの一部，引用情報である。検索語は太字で強調され文脈を知ることができる。検索オプションには，検索条件，著者，出典，日付をそれぞれ指定可能である。出典に「言語処理学会」を指定すると，「自然言語処理」ジャーナルと「言語処理学会年次大会発表論文集」の文献だけが 27 件ヒットし，そのうち 20 件だけ表示される。ここでは NLP 予稿集だけを対象としたいので出典を「言語処理学会年次大会」とした。しかし何もヒットしなかった。

次に J-STAGE で検索した。「コーパス」では 1,966 件ヒットした。詳細検索で「著作権」を追加して本文検索すると 87 件残った。NLP 予稿集に絞る方法を探していると，左コラムに資料名一覧があった。そこには NLP 予稿集はなく，かわりに自然言語処理 (5)の選択肢を見つけた。5 件のヒットである。NLP 予稿集は J-STAGE の対象外であると判断された。ヒット文献の表示様式はグーグルスカラーとほぼ同様であるが，検索語を含むテキストの表示はなく，抄録が表示される。検索語の強調表示はなく，抄録に検索語がない場合も多く，「コーパス」を含むのは 1 件，「著作権」はゼロ。「コーパス」と「著作権」が書かれた文脈を知ることはできず，実際に 1 件ずつ文献を開いて確認したが，同一文脈での用法はなく徒労であった。

コーパス言語資源利用にあたっては著作権への配慮が必要だ。例えば浅原（2017）は，著作物を取り込むのではなく，「検索系を構築し，例文とともに元データが含まれる URL へのリンクを含めて提示するサービスを構築」している。これはグーグルやヤフーなどの検索エンジンと同等の仕組みである。他方，伊藤（2014）は「著作権など法律上の問題も生じると考えられるが，その点については考慮せずまずは理想像を提示する」としている。また藤田，田村（2012）は著作権処理について「本研

究での利用を目的とすることのみについて承諾を得ている．今後，各校と相談した上で公開を検討したい」とした。さらに国立国語研究所による『現代日本語書き言葉均衡コーパス』では「すべてのサンプルについて著作権処理を実施」したと丸山（2010）が紹介している。

3. 目的

文献を探す時，複数の検索語がどのような文脈で使われているかを俯瞰的に示すことにより，利用者が求める情報が書かれた文献に素早くたどり着ける検索エンジンを目指す。その第一弾として NLP 予稿集コーパス専用検索サイト・ナックス（Natural Language Processing Corpus Search Engine; NaCSE; http://www.mintap.com/nacse/nacse.html ）を構築する。

4. 方法

開発手順は以下のとおりであった。

1. NLP サイトのリンクから発表論文（pdf ファイル）収集
2. 収集した pdf ファイルからテキスト（utf-8）抽出
3. NLP サイトの年度ごと論文一覧から，年度・表題・筆者情報を収集
4. 上記情報を統合して検索サイト構築

手順 1 で，第 1 回（1995）から第 23 回（2017）までの 5,260 本を収集した。手順 2 では pdf を txt に変換するソフトを使用した。2003 年以前のものは変換にことごとく失敗した。その結果第 10 回（2004）以降の 14 回分 3,863 本が対象となった。ただし，この中には txt 変換に失敗した文書も 150 本前後見受けられたが，今回は精査しなかった。手順 3 で，論文ごとに 2017_A1-1 のような文献番号を付与した。手順 4 で，公開を前提とし，汎用性を重視して HTML5, javascript を使った。

検索は主検索語と副検索語に区分し，主検索語でヒット状況を表示した後に，改めて副検索語で絞り込む方法を採用した。また，副検索語では「かつ/または」の複合検索を可能とし，文脈理解の利便性を目指した。その他のオプションは公開年のみにした。公開年を採用したのは，検索語が表すテーマの年ごと変動を示すことも大切だからである。

言語資源の著作権は NLP に属している。そこで，浅原（2017）及び他の検索エンジンと同様に検索結果表示に最低限必要な情報を引用しつつ，言語資源本体へのリンクを貼るシステムとした。

5. 結果

図 1 に NaCSE の初期画面を示す。上から順に，検索対象の管理団体名 NLP とそのサイトへのリンク，NaCSE 運営団体，検索エンジンの名称，主検索語入力欄と検索範囲（年），副検索語入力欄と案内を開くボタン，最後が表示切替で文献名だけを表示したり，選んだ文献だけ表示する選択ボタン。

図 1. NaSCE の初期画面

図 2 に検索結果の表示例を示す。例は，主検索が「コーパ

ス」，副が「著作権」，範囲が 2004 年から 2017 年となっている。入力欄のすぐ下から検索結果が表示される。1 行目の「39 件(1825 件中)」は主検索で 1,825 件ヒットし 39 件に絞られたことを示す。下に続くグラフは年ごとのヒット件数。青い棒グラフが主検索結果数，橙が絞り込み結果数。上から下に古い年に遡っている。多少の増減があるものの，コーパスを含む予稿集は 140±40 件程度を推移している。コーパスに関わって著作権に触れているのは毎年 0〜8 件で，ここ 5 年間に目立つ。その下からはヒットした文献が新しい順に並ぶ。同じ年であれば文献番号の若い順になっている。ひとつの文献項目は 4 つのブロックに別れ，上から順に表題（青），文献番号と発表分類（緑），著作者名（緑），ヒットした文脈の引用と続く。表題（青）には文献へのリンクが貼ってある。主検索語「コーパス」は黄色，副検索語「著作権」は橙でハイライトされている。主と副の検索語が一目瞭然でかつ同じ文脈にあることから，検索目的に沿った内容であるかどうかの素早い判断が期待できる。

文献項目の左外にあるスイッチは選択ボタンで，上から 2 番目が選択されている状態である。「チェックだけ表示」をオンにした状態を図 3 に示す。この機能により，読みたい文献を閲覧しやすくしている。2 つ目の文献は，検索語が頻繁に現れていて，探していた文献である可能性が高いことを示唆している。

副検索語は半角セミコロンで「かつ」，半角カンマで「または」の演算子で複合検索可能である。キーワード「著作権,許諾」による複合検索では図 4 のようになる。

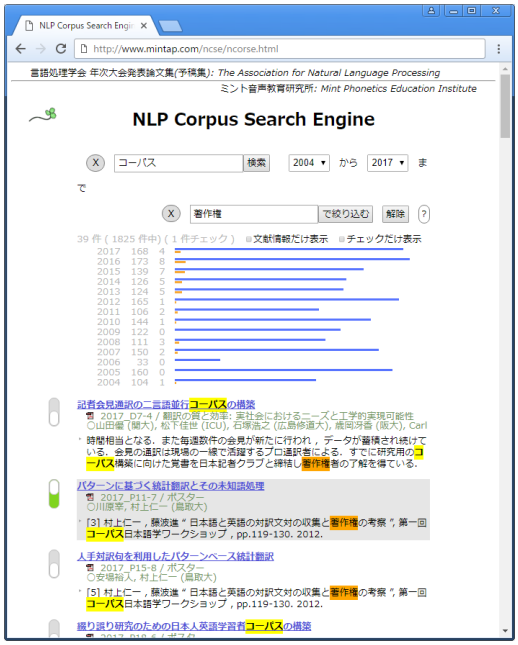

図 2. NaSCE での検索結果表示の様子
主検索語はコーパス、副検索語は著作権

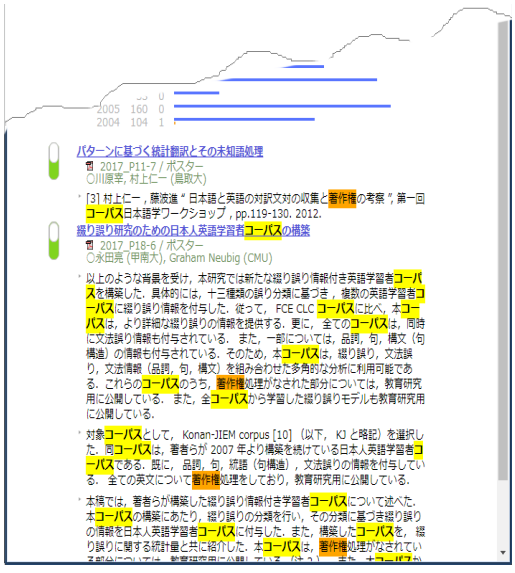

図 3. 検索結果をさらに絞り込んだ様子

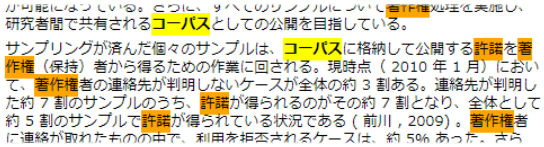
研究者間で共有されるコーパスとしての公開を目指している。
サンプリングが済んだ個々のサンプルは、コーパスに格納して公開する許諾を著作権（保持）者から得るための作業に回される。現時点（ 2010 年 1 月）において、著作権者の連絡先が判明しないケースが全体の約 3 割ある。連絡先が判明した約 7 割のサンプルのうち、許諾が得られるのがその約 7 割となり、全体として約 5 割のサンプルで許諾が得られている状況である（ 前川，2009）。著作権者に連絡が取れた方の中で、利用を拒否されるケースは、約 5%あった。さら

図 4. 「著作権，許諾」による複合検索結果

6. 考察

NaCSE で検索すると，コーパスの先行研究としては，形態素解析（224 件）や語彙（259 件）などが多いが，今回は初級者が所望の文献を文脈検索するエンジン開発をめざした。結果としては，著作権に関する先行研究で取り上げた 4 つの論文を素早く見つける事によって NLP での議論について知ることができるなどの効果を実感した。これまで筆者が使っていたグーグル

スカラーと比べると，効率よく所望の文献にたどり着けた。もしグーグルスカラーが今回開発のNaCSE と同じ表示性能を具備してくれると，より広い範囲での効率のよい結果が期待できる。

7. 議論

今回の開発には幾つかの限界（課題）がある。① 2003 年以前が未収録。② 検索時間が数秒から十数秒かかる。遅い原因は，(1) 検索データの事前整備をしていない，(2) 高速処理のコーディングを採用していない，(3) 一覧性を高める目的で千を超える検索結果でも全表示などだ。③ 対象テキストの改行区切りを単位として検索結果を表示しているが，その単位は表示の 1 行のこともあるので文段落に整序する必要がある。④ 原本の pdf 文書からテキストを抽出した時，2 段組みを検知できずに左右の段組みを無視して横断的につないだ場合がある。③④は膨大な手作業になる。文献執筆時のテキストを利用することができれば③④がすみやかに解決する。

次に著作権に関して述べる。丸山（2010）によるとコーパス構築に関わる著作権処理が必要なサンプル数が 12,604 件あったとされる。これについて「コーパス構築と著作権保護」（前川, 2010）によると，著作権法上の「引用」として処理できるかどうかが議論となったが，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」をしたとある。コーパスの作り方にもよるので単純に一般化できないが，本来「文化の発展に寄与することを目的とする」著作権法が，「国の機関」ほど人も予算も期間も少ない現場に萎縮を生み続けることは好ましいことではないだろう。日本でも「公正な利用であれば, 著作権者の許諾無用」とするフェアユースの考えが広まりつつあることに希望をつなげたい。

8. おわりに

言語処理学会第 24 回年次大会(2018)発表募集の公式サイトにあった「自然言語に関する理論から応用まで幅広い研究発表を募集」し，特に，言語処理とは日頃「縁が薄いと感じておられる人文系」からの「積極的な発表を期待」するとあったので，それに賛同し共感して本コーパスが構築された。

参考文献

浅原 正幸, 岡 照晃（国語研）2017 E1-5 / 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（pp.94-97）

伊藤 薫（京大）2014 P3-2 / 比喩表現コーパスの構築と問題点 -言語学の立場から-（pp.149-152）

藤田 彬, 田村 直良（横浜国大）2012 D4-3 / 作文事例に基づいた児童の「書くこと」に関する学習傾向についての分析–小学四年生による紹介文・感想文を中心に–（pp.987-990）

前川 喜久雄（国語研）「コーパス構築と著作権保護」, Retrieved from http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/Maekawa2010.pdf

丸山 岳彦(国語研) 2010 S1-3 / 代表性を有するコーパスの設計とサンプリングの実際 ―コーパスに基づく言語研究の可能性と限界―（pp.150-153）