「研究ノート」
『パラレルリンク』（Ver.1.0）の開発
ーパラレルコーパス研究の概観とコーパス整備ー

仁科　恭徳・赤瀬川史朗

Abstract

In this paper, we first review previous parallel corpora and analysis studies. We also suggest some future directions in this field. Then, we outline nine parallel corpora included in Parallel Link (Ver.1.0), an online analysis tool for Japanese-English/English-Japanese parallel corpora under development. In particular, we elucidated the text processing, annotation, creation of full-text search indexes, and file organization applied to these parallel corpora.

１．はじめに

本稿では，まず，現在までに構築された日英・英日パラレルコーパス，開発されたコンコーダンサー等の検索ツール，日英・英日パラレルコーパスを活用した研究を概観し，各パラレルコーパスの翻訳方向やテクストジャンル等の特徴と問題点, および今後ニーズが高まるツールや研究について具体的に述べる。次に，これからのパラレルコーパス研究の方向性を示すべく開発中の日英・英日パラレルコーパスオンライン検索ツール『パラレルリンク』（Ver.1.0）に搭載予定の９種のパラレルコーパスの概要と，それらを再整備する上で施したテキスト処理やアノテーション，全文検索インデックスの作成，ファイル整理について詳説する。

２．パラレルコーパス研究の概観

2.1. パラレルコーパス開発研究

現在までに様々な日英・英日パラレルコーパスが開発されている。言語研究で使われてきた古い日英パラレルコーパスの一つに，関西外大コーパスＢ－

日英パラレルコーパス（西村，2002）がある。OS は Windows のみ，専用の検索プログラム Parallel Scan のみで検索可能であったが，2021 年 8 月時点で使用不可である。このコーパスでは，アライメント（例えば，英文 I'm Ken. / 和文「私はケンです。」を，各テクストファイルの同一行番号に配置する処理）が文単位ではなく段落単位となっているため，ParaConc（Barlow, 2002）などで読み込んで使うことはできない。他にも，言語（教育）研究で使われてきた代表的なものに日英新聞記事対応付けデータ（JENAAD）（Utiyama & Isahara, 2003）がある。読売新聞と Daily Yomiuri の対訳データ（一対一対応の日英文，一対多もしくは多対一対応の日英文）で，無料配布は既に終了している[1]。JENAAD を用いた研究には，日・英語間における交換可能性を量的に調査した仁科（2008a）や，英語教育的活用を目的として開発された LWP for ParaNews（公開終了）の授業利用に関する中條他（2015）がある（LWP については次節を参照）。表 1 は現在までに構築された日英・英日パラレルコーパスの例であり[2]，太字の 9 種のパラレルコーパスは，第 3 節で紹介する『パラレルリンク』（Ver.1.0）に搭載予定のコーパスを示す。表 2 はこれら 9 種の対訳対数と語数を示す。

表 1. 2000 年以降に公開された日英・英日パラレルコーパスのまとめ 時系列順（2021 年 9 月現在）

パラレルコーパス	先行研究	翻訳方向	アラインメント単位	ジャンル / レジスター	S/W
Tatoeba 日英対訳コーパス / 田中コーパス（TATOEBA）	Tanaka（2001）, Tatoeba project since 2006	日→英	文	教科書（例：日本人英語学習者が使用している書籍）／歌詞／一般書／聖書の一節	S・W
関西外大コーパス B 日英パラレルコーパス	西村（2002）	日→英	段落	文学／青空文庫と『新潮文庫の 100 冊』に収録されている作品，およびその英訳	W
日英対訳文対応付けデータ（TAIYAKU）	Utiyama & Takahashi（2003）	英→日 日→英（一部）	文	文学 ／ Project Gutenberg, 青空文庫，プロジェクト杉田玄白から 160 作品	W
ロイター日英記事の対応付け（REUTERS）	Utiyama & Isahara（2003）	英→日	文	新聞・ニュース／ロイター通信英・日本語版記事	W
JENAAD 日英新聞記事対応付けデータ	Utiyama & Isahara（2003）	日→英	文	新聞・ニュース／読売新聞，Daily Yomiuri	W
大規模オープンソース日英対訳コーパス（OPENSOURCE）	石坂他（2009）	英→日	文	技術文書／オープンソースソフトウェアのマニュアル	W
Wikipedia 日英京都関連文書対訳コーパス（WIKIPEDIA）	NICT（2010）	日→英	文	ウェブ／ Wikipedia の日本語記事（京都関連）とその英訳	W
TED Talk 日英コーパス（TED）	Cettolo, Girardi & Federico (2012) や https://wit3.fbk.eu/ で初期データ公開	英→日	文単位ではない ※字幕ファイル（ssa 形式ファイル）から作成	アカデミック・ビジネスプレゼンテーション／多種多様なジャンルのプレゼンテーションの字幕データ（音声ファイル付き）	S

日英法令対訳コーパス (LAW)	Neubig（2014）	日→英	段落	法律文書／日本の法律とその英訳	W
SCoRE 用例コーパス（SCoRE）	Chujo, Oghigian & Akasegawa（2015）	英→日	文	教育用例文／教育的に配慮した簡潔で自然な例文（音声ファイル付き）	W
ASPEC（Asian Scientific Paper Excerpt Corpus）	Nakazawa *et al.*（2016）	日→英	文	学術／科学論文の日英アブストラクト	W
Hiragana Times 日英対訳コーパスデータ	2017 年までのマガジン，別冊書籍の日英対訳データ YAC（Your Additional Contact）（https://yac-nippon.com/corpus-english-japanese/en/）	日→英	文（マガジン自体の表示形式は段落単位）	バイリンガルマガジン・書籍／1988 年から 2017 年までの Hiragana Times 349 冊，単行本 19 冊（政治，文化，歴史，恋愛，食べ物，旅行，映画，まんが，習慣など）	W
JESC（Japanese-English Subtitle Corpus）日英サブタイトルコーパス（の一部）	Pryzant *et al.*（2018）	英→日 日→英（一部）	文	映画・ドラマ・テレビ／映画・TV 番組の字幕データ	S

* S/W は Spoken/Written の略である。

表 2. 既存パラレルコーパスの対訳対数と語数（仁科・赤瀬川（2021）を参考。時系列順に改変）

コーパス名	対訳対	語数（日本語）	語数（英語）
TATOEBA	208,013	2,080,831	1,601,860
TAIYAKU	110,909	1,905,586	1,399,650
REUTERS	70,120	2,068,681	1,740,428
OPENSOURCE	505,780	6,927,281	5,018,603
WIKIPEDIA	443,849	9,132,894	9,806,199
TED	518,233	4,657,169	3,247,654
LAW	262,448	9,264,891	9,508,555
SCoRE	10,459	160,337	101,562
JESC	330,102	2,736,837	2,222,329
合計	2,459,913	38,934,507	34,646,840

表 1 から，まず，話し言葉のパラレルコーパスが少なく，最近になって構築され始めたことが分かる。また，この 10 年間でパラレルコーパスはいくつか開発されているものの，自然言語処理・機械翻訳分野で好まれる学術・技術系の専門的なコーパスが多いことも分かる。そして，一部のパラレルコーパスでは翻訳方向が考慮されず，日→英，英→日の双方向翻訳のテクストが混在していることや，アライメントが文単位のものと段落単位のものとに分かれていること，ジャンルに偏りが見られることなども指摘できる。無論，各々に利点があってのテクスト整形ではあるが，これら全てのコーパスを一定ルールのもと統一し合算できれば，ある程度のサイズが保証された擬似的な一般参照日英・英日パラレルコーパスとして言語分析等に活かすことができる[3]。

なお，JENAAD に関しては，2013 年から LWP for ParaNews が公開されオンライン上で簡易検索が可能となっていたが，2021 年 8 月時点で既に公開が終了している（LWP については後節を参照）。また，染谷・赤瀬川・山岡（2011）で使用された Wikipedia 日英京都関連文書対訳コーパス（https://alaginrc.nict.go.jp/WikiCorpus/）も日英パラレルコーパスがレキシカルプロファイラーに実装されている数少ないオンライン検索ツールで，NICT（情報通信研究機構）が 2010 年 10 月に一般公開した Wikipedia の日本語記事（京都関連）とその英訳から構成される日英パラレルコーパスであったが，2011 年 1 月の時点で開発が終了し（最新版は Ver.2.0.1），2021 年 8 月の時点で一般公開はされていない。京都に関する内容が中心で日本の伝統文化，宗教，歴史などの分野をカバーしている。人手翻訳による約 50 万文対を収録し（日本語の語数は約 1,000 万語），翻訳の過程（一次翻訳→流暢さ改善のための二次翻訳→専門用語チェックの 3

段階）を記録している。

表 1 に挙げた以外の日英・英日パラレルコーパスもいくつか存在しており，2021 年 8 月現在リンク切れもあるが日本語対訳データリスト（http://www.phontron.com/japanese-translation-data.php?lang=ja）が参考になるので参照されたい。また，Chujo, Oghigian, & Akasegawa（2015）や Mizumoto & Chujo（2016）など英語教育利用目的のパラレルコーパス検索ツール SCoRE の用例（http://www.score-corpus.org/download/jp/）もダウンロードすることができる。こちらは文単位でアライメントされており，計 10,459 件の英日対応文の用例を獲得することができる。ただし，教育利用が目的である当該コーパスは英文のレベルが統制されている点には留意されたい。

しかしながら，単言語コーパスと比較して日英・英日パラレルコーパスの構築には手間や時間を要することから，現状はその数と種類が限られている。表 1 から現状を把握すると，問題点として，翻訳方向，ジャンル構成比，検索ツール開発の 3 点が挙げられる。特に，今後は翻訳方向（日→英，英→日）ごとに，構成比やバランスも考慮しながら欠落しているジャンルやレジスターのコーパスを追加・構築する必要がある[4]。これは，各特定領域で使用される翻訳ユニットの抽出に有効であるだけでなく，ビジネス文書作成時における実務的利用や，様々な分野における二言語 DDL を用いた教育にも利用できる。そして，このような複数のパラレルコーパスをバランスよく合算することで，擬似的な一般参照パラレルコーパスとしての利用も可能となる。この場合，複数のパラレルコーパスを網羅的に串刺し検索できる使い勝手の良いツール開発も求められる[5]。仮にそのようなツールが開発できれば，二言語辞書に掲載すべき訳語や用例の信頼性と客観性が担保でき（仁科，2008a, 2020），辞書編纂時のコーパス活用の幅も広がる。

2.2. パラレルコーパス検索ツール研究

日英・英日パラレルコーパスの検索ツールには，コンコーダンサーとレキシカルプロファイラーがある。まず，現在使用可能なパラレルコーパス用コンコーダンサーに，Windows のみで使用可能な ParaConc（Barlow, 2002），Windows, Mac，Linux で使用可能な AntPConc（Anthony, 2017），Mac のみで使用可能な CasualPConc（Imao, 2018）がある。このうち，ParaConc のみ有償で，残りの二つは無料で使用できる。CasualPConc の姉妹版 CasualMultiPConc（最新版は Ver.0.4.1）では 2 ～ 5 言語の多言語コーパスの検索・処理が可能である

が，その開発は 2021 年 8 月時点で既に止まっている。前節で挙げた Parallel Scan（西村，2002）も現在使用不可である。

次に，パラレルコーパスが実装され，「見出し語単位で検索，コロケーションなどを文法項目に分類して整理して表示」（染谷・赤瀬川・山岡，2011）することが可能であるレキシカルプロファイラーに LWP（LagoWordProfiler）がある[6]。国立国語研究所と Lago NLP（旧 Lago 言語研究所）が開発したブラウザベースのコーパス検索ツール（パルデシ・赤瀬川，2011）で，単言語コーパスを搭載した代表的なものに，現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese: BCCWJ）の検索を可能とした NLB（NINJAL-LWP for BCCWJ）（パルデシ・赤瀬川，2011）がある。日英パラレルコーパスを搭載したものには，前述の JENAAD の検索が可能な LWP for ParaNews（2021 年 8 月時点で既に公開終了）や Wikipedia 日英京都関連文書対訳コーパスの検索が可能な Wikipedia-Kyoto LWP（WK-LWP）（2021 年 8 月時点で既に公開終了），そして SCoRE（オンライン公開中）がある。なお，Sketch Engine（https://www.sketchengine.eu/）もパラレルコーパスに対応しているが，現時点ではコンコーダンサーとしての機能のみである。

コンコーダンサーの使用には，検索の自由度が高い分，使用者の経験や分析力，直感力等が求められる。一方，レキシカルプロファイラーを使用した場合，あらかじめ決められた文法パターンにそった検索結果が表示されるため，敷居が低い分，検索や分析の柔軟性に欠けるという欠点もある。共に利点・欠点があることから，今後は双方の機能を実装したオンライン検索ツールを開発し無償公開することで一般ユーザーも含めて使用者の増加が見込めるのではないだろうか。特に，様々なジャンル・レジスターのパラレルコーパスをクリック一つで選択でき，検索項目の翻訳ユニットのジャンル構成比なども瞬時に表示できれば，今まで開拓されていなかった研究も可能となろう[7]。

また，京都外国語大学で展開している二言語同時学習（https://www.kufs.ac.jp/faculties/unv_education/unv_program_bi-language.html）をヒントに，開発が止まっている多言語対応コンコーダンサー CasualMultiPConc を用いることができれば，日本語・英語に加え，スペイン語や中国語といった 3 言語以上の複言語 DDL を外国語の授業で展開することもできる。英語に加えて他言語も学んでいる学習者にとって，マルチリンガルコーパスは活用に値するツールとなる。実際に多言語 DDL を教育現場で試した実践例はないため，その有効性を検証すべきであろう。単言語コーパスコンコーダンサーと比較してニー

ズは少ないが，CasualMultiPConc の開発を継続してもらいたい。

2.3. パラレルコーパス活用研究

仁科（2020）でもまとめたが，国内の日英・英日パラレルコーパスを活用した言語記述の研究は多くない。少し古くなるが，2002 年に発刊された『英語コーパス研究』第 9 号に掲載されている 9 本の論文はいずれも当時のパラレルコーパスの構築や検索プログラム，活用事例を知る上で極めて重要である。意味や文法，辞書学的見地から考察したものに，日英再帰形に注目した清水・村田（2002），身体部位を含む日英語表現を分析した岡田（2002），when 節を取り上げた田中（2002）などがある。2002 年以降では，その数は減り，時事英語表現の翻訳傾向などを調査した仁科（2006, 2008a, 2008b），カタカナ語の誤用を取り上げた Nishina（2008），日本語複合動詞とその翻訳を精査した染谷・赤瀬川・山岡（2011），依存木の統語構造的不一致から日英翻訳を分析した Oya（2017），そして，日本語動詞「固める」の翻訳ユニットを日本語コーパス（BCCWJ）と日英パラレルコーパス（WikipediaKyoto LWP）から分析した仁科（2020）などがある（自然言語処理や機械翻訳などの分野では研究報告が目立つ）。なお，これらの中でレキシカルプロファイラーを用いた研究は少なく，染谷・赤瀬川・山岡（2011）と仁科（2020）がそれにあたる。

一方，英語教育関係においては，ツール開発や DDL の教育的活用・効果検証等に関する研究，例えば WebParaNews（https://www.antlabsolutions.com/webparanews/about.html）の中條他（2014, 2015），Anthony, Chujo, & Oghigian（2011）や，SCoRE（http://www.score-corpus.org/）の Chujo, Oghigian, & Akasegawa（2015），Mizumoto & Chujo（2016）などの論文が発表されている。特に SCoRE に関しては，「慎重に作成した簡潔で自然な英語例文約 10,000 文と，日本人英語教師が丁寧に付けた日本語対訳文」から構成され，英文レベルが統制されている。日・英語の記述を分析する場合は，収録された英語テクストが制限・統制なしで産出されたものが理想であるため，言語研究と教育研究の利用とで求められるパラレルコーパスの質が異なることに留意されたい。

これからのパラレルコーパス研究に関しては，複数のパラレルコーパスが予め搭載され一括検索できる検索ツールの開発が進めば，日英・英日翻訳の量的分析が容易となり，特定の語・句のジャンルごとの翻訳実態の解明や，英和・和英辞書あるいは辞書データベースに掲載されている訳語・訳例を客観的に精

査することが可能となる。あるいは，一般的に我々が想定しているものとは異なる質のパラレルコーパスを用いた研究も考えられる。例えば，複数の翻訳家による翻訳ストラテジーの計量的調査を見込んだ一対多のパラレルコーパスの構築とその研究も興味深い。具体的には，同一の起点言語のテクスト（source text）と複数の翻訳家によって作成されたその目標言語のテクスト（target text）の各文をアライメントすることで構築される一対多のパラレルコーパスを活用すれば，各翻訳家の翻訳ストラテジーの類似性や相違性が可視化できる。熟達した翻訳家が暗黙に共有している翻訳ストラテジーの解明のみならず，各翻訳家の個性もデータとして可視化できるのである。参考までに，ルイス・キャロル著 *Alice's Adventures in Wonderland* には翻訳家 39 人による計 55 の日本語訳版が存在し，*Through the Looking-Glass* には 20 人の翻訳家による計 27 の日本語訳版が存在している（詳しくは，http://www.hp-alice.com/lcj/g_contents.html）。また，楠本（2001: 4-7）によれば，1998 年時点で両アリスの作品は 150 種前後が存在しているという説もある。これらの翻訳作品で用いられた表現や翻訳手法を計量的に比較するためには一対多のパラレルコーパスの構築が必要不可欠であり，筆者の知るところ，現時点でそのような研究は皆無である。今後の翻訳研究を前進させる上でその構築と分析には一定の価値があろう。

3．『パラレルリンク』（Ver.1.0）の開発準備

前節までの日英・英日パラレルコーパス（研究）の状況を受け，既存の日英・英日パラレルコーパスを串刺し検索できるオンライン検索ツール『パラレルリンク』（Ver.1.0）を Lago NLP（旧 Lago 言語研究所）と共同開発中である。仁科・赤瀬川（2021）が示すように，Ver.3.0 までを予定している約 10 年計画の本プロジェクトでは，最終的に各ジャンルにつき双方向翻訳のパラレルコーパスを搭載し，欠落しているジャンルのコーパスについては一からの構築も検討している。本節では，その第一段階として取り組んでいるプロトタイプに搭載予定のパラレルコーパスの選定と，それらのテクスト処理，アノテーション，全文検索インデックスの作成，ファイル整理について説明する。

3.1.『パラレルリンク』（Ver.1.0）に搭載予定のパラレルコーパス

本ツールの開発に先立ち，既存のパラレルコーパスの中身を再整備した。対象とした日英・英日パラレルコーパスは，表 1 中において太字で示した計 9 種

である[8]。大規模ウェブパラレルコーパス JParaCrawl（1,000 万対）（http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/）も収録を検討したが，ノイズが多いため今回は見送った。また，アカデミック分野のパラレルコーパス Asian Scientific Paper Excerpt Corpus（ASPEC）（300 万対）（http://orchid.kuee.kyoto-u.ac.jp/ASPEC/）の収録も検討したが，今回の『パラレルリンク』は一般公開を目指していることから，研究利用に限定されている当該コーパスは含めていない[9]。一方で，仁科（2020）では映画やドラマの日英・英日字幕コーパスを含める有用性に触れたが，今回，ノイズが多いものの JESC は含める方針を採った。また，Tatoeba コーパスに収録されている英和対訳文の元になった田中コーパスは，学生が翻訳した対訳文を数年かけて収集した約 15 万対のコーパスであるが，会話文が全体の 40% を占めることから含めることにした[10]。

なお，教育目的では SCoRE 用例コーパスを活用することが最適であるが，言語学的な分析に関しては今回選定した 9 種の中では，SCoRE を除く他 8 種のパラレルコーパスを用いるべきかもしれない（仁科，2020 参照）。また，各コーパスサイズが異なることから，コーパス間で比較分析（ジャンル・レジスター分析）を可能にするために，コーパスごとに 100 万語あたりの生起頻度を表示する機能をインターフェースに実装する予定である。しかしながら，今回搭載した 9 種のコーパスだけでは検索したい語・句の十分な翻訳例が得られない可能性もあるため，表 1 に挙げた他のパラレルコーパスや自作コーパスの追加も Ver.2.0 以降に検討したい[11]。

3.2.『パラレルリンク』（Ver.1.0）のテクスト処理・アノテーション

次に，各パラレルコーパスのフォーマットを統一するためにテクスト処理を施し，英語には品詞情報，日本語には形態素情報を付与した。そして，Blacklab Query Tool（https://inl.github.io/BlackLab/query-tool.html）を用いて全文検索のインデックスを作成した。まず，テクスト処理としてテクストのクリーニング，エンコーディングの統一（UTF8），フォーマットの統一，センテンス ID の付与を施した。以下は，対訳ファイルのサンプル（TED）である。次に，品詞情報・形態論情報を付与した。まず，英文に関しては，Stanford POS Tagger（https://nlp.stanford.edu/software/tagger.shtml）を用いて，表層形，レマ，品詞など品詞に関する情報を付与した。また，日本語に関しては形態素解析器 Sudachi（https://github.com/WorksApplications/

Sudachi）を使用し，表層形，語彙素，品詞に関する形態論情報を付与した。今後，文単位への変換を予定している。

```
TED 00001 0000000001 I'm going to talk to you tonight 今晩 お話しするのは
TED 00001 0000000002 about coming out of the closet カミングアウトについてです
TED 00001 0000000003 and not in the traditional sense いわゆる「カミンアウト」
TED 00001 0000000004 not just the gay closet. ゲイだと打ち明けることではありません
TED 00001 0000000005 I think we all have closets. 誰しも心に壁を作っています
TED 00001 0000000006 Your closet may be telling someone その後ろに隠れているのは
TED 00001 0000000007 you love her for the first time 誰かに初めて愛の告白をすることや
TED 00001 0000000008 or telling someone that you're pregnant 妊娠したこと
TED 00001 0000000009 or telling someone you have cancer ガンであることを伝えることかもしれません
TED 00001 0000000010 or any of the other hard conversations 他にも私たちが人生で経験する一
```

図 1. 対訳ファイルサンプル（TED）

3.3.『パラレルリンク』（Ver.1.0）の全文検索インデックスの作成

その後，全文検索インデックスを作成した。詳しくは，上記の品詞情報，形態論情報を含む Blacklab Query Tool のインポートファイルを作成した。ファイル形式は XML である。

```
<?xml version="1.0" encoding="UTF-8"?>
<docs>
 <doc corpus="TED" subcorpus="" fid="00001" sid="0000000001" type="en" counterpart="
今晩 お話しするのは ">
  <s id="TED:;00001:0000000001">
   <w p="PRP" l="I">I</w>
   <w p="VBP" l="be">'m</w>
   <w p="VBG" l="go">going</w>
   <w p="TO" l="to">to</w>
   <w p="VB" l="talk">talk</w>
   <w p="TO" l="to">to</w>
   <w p="PRP" l="you">you</w>
   <w p="RB" l="tonight">tonight</w>
  </s>
 </doc>
```

図 2. インポートファイルサンプル（TED 英文）

```
<?xml version="1.0" encoding="UTF-8"?>
<docs>
  <doc corpus="TED" subcorpus="" fid="00001" sid="0000000001" type="ja"
counterpart="I&#x27;m going to talk to you tonight">
    <s id="TED:;00001;0000000001" corpus="TED" type="ja">
      <w p=" 名詞 , 副詞可能 ,*,*" l=" 今晩 "> 今晩 </w>
      <pu> </pu>
      <w p=" 名詞 , サ変接続 ,*,*" l=" お話し "> お話し </w>
      <w p=" 動詞 , 自立 ,*,*" l=" する "> する </w>
      <w p=" 名詞 , 非自立 , 一般 ,*" l=" の "> の </w>
      <w p=" 助詞 , 係助詞 ,*,*" l=" は "> は </w>
    </s>
  </doc>
```

図 3. インポートファイルサンプル（TED 日本語文）

3.4. ファイル整理

処理後のテクストファイルについては，大きく分けて 3 種のフォルダ（formatted; annotated; blacklab）に整理した。まず，formatted フォルダには，パラレルコーパスの種類ごとに 9 種類のサブフォルダ（JESC; Law; OpenSource; Reuters; SCoRE; Taiyaku; Tatoeba; TED; Wikipedia）が用意されている。また，各サブフォルダには，それぞれ以下 2 種類のテクストファイルが収録されている。コーパスデータは，コーパス，サブコーパス，ファイル ID，センテンス ID，英文，和文の 6 つのフィールドから構成され，各フィールドはタブで区切られている。エンコーディングは前述のとおり，UTF-8 で統一している。リンクデータは，コーパス，サブコーパス，ファイル ID，ファイル名の 4 つのフィールドから構成され，各フィールドはタブで区切られている。検索ツールを開発するときに，元のコーパスファイルを表示するために用いられる。また，original サブフォルダも用意し，こちらには変換前の元データが収録されている。

```
[ コーパス名 ].txt... 統一フォーマットのコーパスデータ
[ コーパス名 ].metadata.txt... 元のコーパスファイルとファイル ID とのリンクデータ
```

図 4. 各サブフォルダに収録されている 2 種類のテクストファイル

次に，annotated フォルダには，formatted フォルダにある統一フォーマットのコーパスデータにアノテーション情報を付与したファイルを収納している。ファイルのフォーマットは前述のとおり XML ファイルで，Blacklab Query Tool のインポートファイルとなる。各コーパスについて，英文と和文の 2 種類の XML ファイルが用意されている。

最後に blacklab フォルダには，Blacklab Query Tool のインデックスファイルが収録されている。検索ツールのバックエンドの役割を果たす。以上のようなテクスト処理，アノテーション，全文インデックス作成，ファイル整理を行うことで，既存パラレルコーパスを整備した。

4. まとめ

本稿では，はじめに過去から現在までの日英・英日パラレルコーパスや検索ツール，それらを活用した研究の変遷を振り返り，これからの展望を述べた。そして，現在開発中の日英・英日パラレルコーパスオンライン検索ツール『パラレルリンク』（Ver.1.0）に搭載予定のパラレルコーパスの概要とテクスト処理などの一連の再整備作業について詳説した。当該検索ツールのインターフェースや実装している検索機能の紹介，活用研究などについては，論を改める。

謝辞

本研究は JSPS 科研費 20K00692 の助成を受けたものである。また，2021 年 10 月 2 日にオンラインにて開催された英語コーパス学会第 47 回大会において口頭発表したものに，大幅な加筆・修正を施したものである。ここに，『パラレルリンク』の開発に携わって頂いた第二著者の Lago NLP（旧 Lago 言語研究所）の赤瀬川史朗代表，および SCoRE の用例コーパスを搭載することをご快諾くださった中條清美先生（元日本大学），現在 SCoRE の一連の研究を引き継いでおられる西垣知佳子先生（千葉大学），ならびに関係者の皆様に感謝の意を示す。

注

1. 詳しくは，http://www2.nict.go.jp/astrec-att/member/mutiyama/index-ja.html を参照。

2. カーネギーメロン大学の Graham Neubig 氏のウェブサイト http://www.phontron.com/japanese-translation-data.php?lang=ja も参考にした。
3. 一般参照と呼ぶには，サブコーパスのサイズやバランス，構築デザインそのものを統制することが難しいため，「擬似的な」ということばをここでは用いた。
4. 翻訳物ということを考えると，収集テクストの時代性を統一することが難しく，原著のみならず翻訳物の版権の問題もあるため，そのハードルは想像以上に高い。
5. 既存のパラレルコーパスを再整備して一覧検索を可能にすることが，現時点でできる第一歩であろう。これが，第 3 節で紹介する『パラレルリンク』（Ver.1.0）の開発へと繋がっている。
6. 赤瀬川・パルデシ・今井（2014, p.41）によれば，レキシカルプロファイリングとは「あらかじめ設定された検索式に基づいて，コーパスから様々なタイプのコロケーションの情報を抽出した結果を，文法パターンごとに整理してユーザに提示するコーパス検索手法」であり，「特定の語彙の文法的振る舞いやコロケーションをマクロ的視点から調査できる」と説明する。
7. 第 3 節で紹介する『パラレルリンク』（Ver.1.0）では，手始めにレキシカルプロファイラーの機能を実装する予定であるが，Ver.2.0 以降ではコンコーダンサーも搭載する予定である。
8. コーパス・デザインから構築，版権取得までかなりのハードルがあるため，単言語コーパスと異なりパラレルコーパスの新たな開発や普及はそれほど進んでいない。よって，現時点で何ができるかを考えた場合，既存資源を有効活用する『パラレルリンク』には一定の意味があろう。
9. 内部利用はおそらく可能であることから，使用者を限定した研究者用のツール開発も進めたい。
10. 正確には 146,784 文が日本語と英語の両方で書かれており，大半が短文であり，英文の長さが平均で 7.72 語，最長で 45 語との報告がある（http://hihan.hatenablog.com/entry/2019/01/20/070254）。また，学生 1 人あたり 300 個の文章を翻訳したことから，翻訳者が多数存在する一方で複数の日本人大学生が翻訳プロジェクトに参加したため誤訳が混ざっている可能性もあり，質の点では保証できないという欠点がある。
11. ただし，JENAAD は既に無償配布が終了しているため，使用許諾に費用が発生する（JENAAD の有償ライセンスは非商用で 50 万程度である）。同様に Hiragana Times 日英対訳コーパスデータのアカデミックユースは一般の 40% 引きの 150 万程度で契約可能である。

参考文献

赤瀬川史朗・パルデシプラシャント・今井新悟（2014）「NINJAL-LWP の類義語比較機能」『第 6 回コーパス日本語学ワークショップ予稿集』41-50.

Anthony, L. (2017) AntPConc (Ver.1.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.html

Anthony, L., K. Chujo and K. Oghigian (2011) "A Novel, Web-based, Parallel Concordancer

for Use in the ESL/EFL Classroom." In Newman, J., H. Baayen and S. Rice (eds.), *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. New York: Rodopi, pp. 123-138.

Barlow, M. (2002) ParaConc: Concordance Software for Multilingual Parallel Corpora. [Computer Software]. Available from https://paraconc.com

Cettolo, M., C. Girardi and M. Federico. (2012). "WIT3: Web Inventory of Transcribed and Translated Talks." *Proceedings of the 16th EAMT Conference, 28-30 May 2012*: 261-268.

中條清美・アントニローレンス・内山将夫・西垣知佳子（2014）「フリーウェア WebParaNews オンライン・コンコーダンサーの英語授業における活用」『日本大学生産工学部研究報告 B』第 47 号：49-63.

中條清美・西垣知佳子・赤瀬川史朗・内山将夫（2015）「レキシカル・プロファイリング型オンラインコーパス検索ツール LWP for ParaNews の英語授業における利用」『日本大学生産工学部研究報告 B』第 48 号：45-57.

Chujo, K., K. Oghigian and S. Akasegawa (2015) "A Corpus and Grammatical Browsing System for Remedial EFL Learners." In Leńko-Szymańska, A., and A. Boulton (eds.), *Multiple Affordances of Language Corpora for Data-driven Learning*. Amsterdam: John Benjamins, pp. 109-128.

Imao, Y. (2018) CasualPConc (Ver.1.0) [Computer Software]. Available from https://sites.google.com/site/casualconcj/ その他のアプリケーション /casualpconc

石坂達也・内山将夫・隅田英一郎・山本和英（2009）「大規模オープンソース日英対訳コーパスの構築」『情報処理学会研究報告』第 17 号：1-7.

楠本君恵（2001）『翻訳の国の「アリス」－ルイス･キャロル翻訳史･翻訳論』未知谷 .

Mizumoto, A., and K. Chujo (2016) "Who is Data-driven Learning for? Challenging the Monolithic View of its Relationship with Learning Styles." *System 61*: 55-64.

Nakazawa, T., M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi and H. Isahara (2016) "ASPECT: Asian Scientific Paper Excerpt Corpus." *Proceedings of the 9th International Conference on Language Resources and Evaluation*: 2204-2208.

Neubig, G.（2014）日英法令対訳コーパス . Available from http://www.phontron.com/jaen-law/index-ja.html

NICT（2010）Wikipedia 日英京都関連文書対訳コーパス（Ver.2.0.1）. Available from https://alaginrc.nict.go.jp/WikiCorpus/

西村公正（2002）「誌上シンポジウム 日英パラレルコーパスでどのような英語研究が可能か -- コーパス構築の概要と検索プログラム , および研究事例」『英語コーパス研究』第 9 号：37-43.

仁科恭徳（2006）「相互関係を表す形容詞から見たシノニム学習の理論と実践教材 : 実証的考察とパラレルコーパスを用いたデータ駆動型学習法を中心に」『LET 関西支部研究集録』第 11 号：45-59.

仁科恭徳（2008a）「パラレルコーパスを用いた交換可能性の一考察」『英語コーパス研究』第 15 号：81-95.

仁科恭徳（2008b）「パラレルコーパスを用いた抽象語彙・フレーズの一考察 : これからの二言語辞書の編纂論」『LET 関西支部研究集録』第 12 号：83-97.

Nishina, Y. (2008) "Parallel Corpora in Computer-assisted Language Learning: A Case of Lexical Studies and Data-driven Learning Using Moodle". In Marriott, R., and P. Torres (eds.), *Handbook of Research on E-Learning Methodologies for Language Acquisition.* Hershey: Information Science, pp. 203-217.

仁科恭徳（2020）「日英パラレルコーパス WikipediaKyoto-LWP を用いた和英辞典の記述改善案について－「X を固める」の場合－」『英語コーパス研究』第 27 号：1-21.

仁科恭徳･赤瀬川史朗（2021）「日英･英日パラレルコーパスオンライン検索ツール『（仮称）パラレルリンク』（Ver.1.0）の開発に向けて（中間報告）」『英語コーパス学会大会予稿集 2021』25-30.

岡田啓（2002）「「顔」を含む日本語表現と対応する英語表現について」『英語コーパス研究』第 9 号：57-79.

Oya, M. (2017) "Syntactic Divergence Patterns among English Translations of Japanese One-word Sentences in a Parallel Corpus." *English Corpus Studies 24*: 19-40.

パルデシプラシャント・赤瀬川史朗（2011）「BCCWJ を活用した基本動詞ハンドブック作成－コーパスブラウジングシステム NINJAL-LWP の特長と機能－」『現代日本語書き言葉均衡コーパス完成記念講演会予稿集』国立国語研究所，pp. 205-216.

Pryzant, R., Y. Chung, D. Jurafsky and D. Britz (2018) *JESC: Japanese-English Subtitle Corpus*. Ithaca, New York: Cornell University. Available from https://arxiv.org/pdf/1710.10639

清水眞・村田真樹（2002）「パラレルコーパスを用いた日英再帰形の分析」『英語コーパス研究』第 9 号：17-34.

染谷泰正・赤瀬川史朗・山岡洋一（2011）「大規模翻訳コーパスの構築とその研究および教育上の可能性」『日本メディア英語学会第 1 回年次大会発表資料』1-15.

田中美和子（2002）「『語り』の when 節」の意味特徴」『英語コーパス研究』第 9 号：81-91.

Tanaka, Y. (2001) "Compilation of a Multilingual Parallel Corpus." *Proceedings of PACLING 2001*: 265-268.

Utiyama, M., and H. Isahara (2003) "Reliable Measures for Aligning Japanese-English News Articles and Sentences." *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics 2003*: 72-79.

Utiyama, M., and M. Takahashi (2003) *English-Japanese Translation Alignment Data*. Available from https://www2.nict.go.jp/astrec-att/member/mutiyama/align/index.html

（仁科　恭徳　神戸学院大学　Email: ynishina@gc.kobegakuin.ac.jp）

（赤瀬川史朗　Lago NLP（旧 Lago 言語研究所）Email: shiro.akasegawa@lagonlp.jp）