日本語と英語の対訳文対の収集と著作権の考察

村上仁一 (鳥取大学 工学部 知能情報工学科) [*1]
藤波進　（学際統合創研株式会社） [*2]

Japanese-English Parallel Sentences Collection from Digital Media

Jin'ichi Murakami (Faculty of Engineering, Tottori University)
Susumu Fujinami (Cyber Creative Institute CO. Ltd)

概要

日英対訳辞書は，翻訳の研究において必要不可欠のものである．しかし，日本文と英文が文単位に対応していて，量が多く，一般の人が入手可能な対訳文対は，最近まで存在していなかったと言える．本研究では，様々な電子媒体から，日本文と英文が文単位に対応する対訳文対を採取した．電子媒体として，CD-ROM・Internet・電子辞書などを利用した．これらの結果，対訳文対として 1,099,093 万文対を採取した（対訳データベース）．そして，得られた対訳文対から，日本文において単文の対訳文対を 182,113 文対，採取した（単文データベース）．また日本文において重文・複文の対訳文対を 158,633 文対，採取した（重文・複文データベース）．また最後に著作権の問題について述べる

キーワード：機械翻訳，統計翻訳，日英対訳文対，単文，重文・複文，著作権

Summary

We collected large number of Japanese-English parallel sentences from many digital medias. There are many digital media like Japanese English dictionaries, English sample sentences and CD-ROMs. There are 9 types in digital medias. Finally, We collected about 1,099,093 parallel sentences. Also, we extracted 182,113 simple sentences from this parallel sentences. And, we extracted 158,633 complex and compound sentences from this parallel sentences. Also, we described the copyright problem for parallel sentences.

Key words : Machine Translation, Statistical Machine Translation, Parallel Corpus, Simple Sentence, Complex Sentence, Compound Sentence, Copy right

1　はじめに

日英・英日対訳辞書は，翻訳において必要不可欠のものである．そのため，辞書編纂には長い歴史がある．しかし，つい最近まで，電子的に読めて，一般の人が入手可能であり，大規模で，日本文と英文が文単位で対応している対訳文対は，存在していなかったと言える．現在，多くの英日や日英の辞書や例文集などが CD-ROM などの電子媒体で販売されている．しかし，日本文と英文が文単位で対応している対訳文対を大量に採取することは，困難である．電子媒体をある程度加工することで対訳文対を作成することが可能である．しかし，対訳文対を採取することが困難なコーパスは多い．

本研究では，様々な電子媒体のコーパスから得られた日英対訳文対の量や抽出の問題点や得られた対訳文対について述べる．電子媒体には，電子辞書・CD-ROM・Internet などがあるが，本研究では，以下の 9 つに分類した．

1 電子辞書，共通フォーマット　2 電子辞書，独自フォーマット　3 CD-ROM 付の書籍　4 Internet　5 新聞記事
6 対訳文対，無償　7 対訳文対，販売　8 対訳文対，未販売　9 その他

これらの電子媒体から，本研究では，総計として，1,099,093 文対の対訳文対を採取した．そして，得られた対訳文対から，日本語において単文の対訳文対を 182,113 文対，採取した．また日本語において重文・複文の対訳文対を 158,633 文対，採取した．

最後に，これらの対訳文対における著作権の問題について述べる．

2　利用可能な電子媒体

現在，多くの英日や日英の辞書や例文集が，CD-ROM などの電子媒体で販売されている．しかし，日本文と英文が文単位で対応している対訳文対をもつ電子媒体は少ない．しかし，電子媒体をある程度加工することで，対訳文対を採取することが可能である．採取可能な電子媒体は，以下の 9 つに大きく分類することができる．以後は，それぞれの特徴について述べる．

1. 電子辞書，共通フォーマット （分類番号 1）

   現在，コンピュータにおいて検索可能な日英，英日辞書の電子辞書がある．このフォーマットには，一般に公開されているフォーマットを使用している辞書と，各社独自のフォーマットを採用している辞書の 2 種類がある．

   一般に公開されているフォーマットには EPWING 形式と電子ブック形式とロボワード形式の 3 種類が有名である． EPWING は日本独自の電子出版の共通フォーマットで，基本的に JIS コードで記録されている．このフォーマットの辞書は，英日，日英辞書のほかにも広辞苑や漢和辞書などがあり，現在 50 種類を越える辞書が販売されている [41]．フォーマットが公開されているため，テキストを抽出することは容易である．しかし，対訳文対の採取は，かなり困難な場合が多い．通常の辞書では，掲載されている文がそのままテキストになっている．つまり，例文はテキストの中に埋め込まれている．そのため対訳文対は特定の記号や空白などを手がかりにして採取する必要がある．したがって対訳文対の採取は辞書ごとに異なるプログラムを書く必要がある．

   なお，例外として対訳文対が用例ファイルになっている辞書がある．例として，斎藤英和大辞典（3.13 節）があげられる．このような辞書は，簡単に対訳文対を採取することが可能である．しかし種類は少ない．

2. 電子辞書，独自フォーマット （分類番号 2）

   電子辞書では独自のフォーマットをとり，専用のブラウザでなければ見えないものがある．これらの解析は非常に手間のかかる．特に辞書に圧縮がされている場合や外字がある場合，解析は困難である．しかし，ランダムハウス英語辞典（3.17 節）は，歴史のある辞書であるため，フォーマットを解析して EPWING 形式に変換するツールが Internet 上で掲載されている．また，ビジネス技術実用英語大辞典（3.18 節）は解析可能であった．類似の辞書は他にもあるが，解析に時間がかかるため，対訳文対の採取は行わなかった．

3. CD-ROM 付の書籍 （分類番号 3）

   最近は CD-ROM 付の書籍が販売されている． この中から，日英の対訳のある書籍を選んで，簡単なスクリプトをつくることで，対訳文対が採取できる．ただし，1 冊において得られる対訳文対は少ない．例として，“英文ビジネスレター実用フォーマット” （3.35 節）と “英文 E メール文例集”（3.36 節）がある．

[*1] murakami@ike.tottori-u.ac.jp
[*2] http://www.cybersoken.com/

4. Internet （分類番号 4）
   Internet 上に公開されている対訳文対がある．基本的には，中学校や高校における英文法の教育用の定型文である．代表的な例としてアルク社がある． ただし，一定の時期にしか公開されていない．対訳文対を採取した例として，英語教師用データベース（3.24 節）がある．
5. 新聞記事 （分類番号 5）
   大手の新聞社では，日本語の記事と英語の記事が同時に発行されている．朝日新聞と Asahi Evening News，読売新聞と The Daily Yomiuri，毎日新聞と Mainichi Daily News がある．これらは個別に CD-ROM で購入できる．しかし，対訳文対の採取は，原文が記事対応にすらなっていないため，非常に困難である．しかし，日本文と英文の対訳文対を自動的に採取する研究があり，これを利用して，約 20 万文対の対訳文対が作成されている [Utiyama (2003)]．ただし記事対応のテキストから自動的に対訳文対を採取しているため，他のコーパスから採取した対訳文対と比較すると英文の誤りが多い（5.1 節）．また，元の新聞記事の CD-ROM の販売価格は非常に高価である．
6. 対訳文対，無償 （分類番号 6）
   無償で配布している対訳文対がある．多くは個人が収集したコーパスである．研究用には，自由に使用可能と思われる．例としては，田中コーパス (3.32 節) がある．この対訳文対は，元兵庫大学の故田中康仁氏が収集したものである．単文が多くを占めていて，約 20 万文対ある．ただし，学生が英文を作成したため，英文の品質は低い．
7. 対訳文対，販売 （分類番号 7）
   わずかな例ではあるが，英日の対訳文対として販売されているコーパスがある．研究用には，自由に使用可能と思われる．例としては，英文ビジネスレター文例大辞典 (3.8 節) がある．
8. 対訳文対，未販売 （分類番号 8）
   基本的には，個人もしくは会社が翻訳の研究のために収集したコーパスである．翻訳の研究のために作成されたものであるため，対訳文対として最適な文対が得られる．しかし，残念なことに一般の人には入手不可能である．例として IPAL （3.2 節）がある．この分類に属するコーパスの多くは，故池原悟氏が電電公社および NTT に在籍しているときに作成した対訳文対である．
9. その他 （分類番号 9）
   現在，大規模な日英対訳コーパスが公開され始めている．一方で公開されていないコーパスもある．以下に，これらの例を挙げる．

（a）特許
日本文と英文の特許文から対訳文対を採取し，これを利用して翻訳のコンテストが開催されている [Fujii (2010)]．利用可能な対訳文対は 500 万文対を超える．現在入手可能な日英の文単位の対訳文対として最大であろう．ただし，特許文であるため通常の日本文とかなり異なる．また，同一の日本文と英文の特許から，プログラムで自動的に対訳文対を作成しているため [Utiyama (2003)]，誤った対訳文対がある．

（b）旅行対話タスク
旧 ATR 現 NICT では，旅行文コーパス（BTEC : Basic Travel Expression Corpus）を収集している．全部で約 70 万文対あると思われる．残念ながら，一般に公開されていない．しかし，IWSLT のコンテストに参加した場合，約 2 万文対が利用可能である．

（c）Wikipedia 日英京都関連文書対訳コーパス
NICT の MASTAR プロジェクトにおいて，2010 年に『Wikipedia 日英京都関連文書対訳コーパス Version 2.0』が公開された．Wikipedia を翻訳したもので，合計約 50 万文対ある．

（d）青空文庫とプロジェクト杉田玄白
現在 Internet 上において，著作権が切れた本を掲載する青空文庫が発足している [42]．これらの本に対して翻訳をする杉田玄白プロジェクト [43] がある．このプロジェクトを利用することで対訳文が得られるが，残念ながら文単位の対訳文対になっていない．

（e）みんなの翻訳
「みんなの翻訳」は，翻訳文を共有することで，翻訳の効率改善と発展を目的として発足した [44]．ただし，プロジェクト杉田玄白と同様，文単位の対訳文対になっていない．

（f）白書
日本国憲法や政府機関が発行する白書（教育白書など）を英文に翻訳したテキストがある．ただし，訳文は，文単位の対訳文対になっていない．

（g）NHK
NHK はニュースなどの 2 カ国語放送があるため，大量の対訳文対がある．しかし，放送法による制約のため外部への提供はできないと思われる．

（h）外国語大学
各外国語大学には，独自で収集した対訳文対を持っている教官がいるが，全容は不明である．

（i）日本語 WordNet
日本語 WordNet[45] には 48,276 の例文が掲載されている．ただし，句と文が入り交じってる．また文頭が大文字になっていないなどの問題点がある．

（j）EDICT
EDICT[46] には例文が掲載されている．ただし，日本文と英文の対訳文対に変更するのは，やや困難である．

3 利用したコーパス

以下に本論文で利用した各コーパスの概略を述べる． なお，アンダーラインで囲まれたイタリックの文章は，各コーパスの販売用の案内文や紹介文の抜粋である．なお節の角括弧の番号は参考論文の番号である．また節の丸括弧のアルファベットは表 11 および表 12 の ID 番号である．

3.1 機能試験文集 [1](AA,BC)

分類は 8．一部は分類 4． 故池原悟氏が NTT に在籍のときに作成し，機械翻訳システム評価を目的とした対訳文対である． オリジナルは 5,240 文対であるが，公開されている文対は 3,718 文対である．現在は，“http://www.kecl.ntt.co.jp/mtg/resources/index-j.php” にて公開されている．なお，表 11，表 12 において “機能試験文集（公開）” は公開されている 3,718 文対である．

3.2 IPAL[2] (AB)

分類は 8. 原典は，情報処理振興事業協会 (IPA) が作成した日本語の辞書である．IPAL の辞書の一部の『計算機用日本語基本動詞辞書 IPAL』『計算機用日本語基本形容詞辞書 IPAL』『計算機用日本語基本名詞辞書 IPAL』に含まれる日本文を，NTT の研究所で英文に翻訳した．

なお，現在 IPAL は，GSK[言語資源協会] において無料配布されている．

3.3 学研 アンカー和英辞典，アンカー英和辞典 [3] (AC,AD)

分類は 1．英和約 51,000 語を収録した『ニューアンカー英和辞典』，和英約 25,000 語を収集した『ニューアンカー和英辞典』の CD-ROM である．

3.4 学研 英和辞典 (AE)

分類は 8. 学研が，アンカー和英英和辞書を発売する前に販売していた辞書である．この辞書は現在販売されていない．この辞書の例文を，NTT の研究所が人手で入力した対訳文対が存在する．

3.5 外国人のための基本語用例辞典 [4](AF)

分類は 8. 文化庁から出版されている “外国人のための基本語用例辞典” の日本文を，MTT の研究所で英文に翻訳した．鳥取県岩美町の澤田信一氏が訳したとのデータがある．

3.6 三省堂 英語表現辞典 [5](AG)

分類は 8. 英語表現辞典の中の例文を，MTT の研究所が人手で入力した対訳文対である．

3.7 日本経済新聞 (AH)

分類は 8. 日本経済新聞の日本文を，NTT の研究所で英文に翻訳したコーパスである．新聞記事のため，日本文が非常に長い．そのため英文も長い．

3.8 英文ビジネスレター文例大辞典 [6](AI)

分類は 7. 文単位の英日対訳文対として販売されている，希少なコーパスである．ただし，現在販売が中止されていると思われる．

3.9 外国人のための日本語例文・問題シリーズ [7](AJ)

分類は 8. 外国人のための日本語例文・問題シリーズの中の日本文を，NTT の研究所で翻訳した．助詞，語彙，形式名詞，表記法，談話の構造などの編に分かれている．

3.10 自然発話音声・言語データベース (LDB)[8] (AK)

分類は 7. ATR が作成した．自動音声の研究のためにホテル対話などが収録されている．コーパスには音声とテキスト両者が含まれる．このコーパスには，約 20 万文対の対訳文対がある．

3.11 SENSEVAL 対訳コーパス [9] (AL)

分類は 4. senseval は，語の意味的曖昧性解消のためのコンテストである．この日本語タスクには，辞書タスクと翻訳タスクがあり，翻訳タスクでは，日本語単語に対する適切な英訳を選択することを目指している．このタスクのために対訳コーパスが作成されている．ただし，このコーパスの多くは単語の訳であり，対訳文対は少数である．

3.12 講談社和英辞典 [10] (AM)

分類は 6. 講談社和英辞典から電子技術総合研究所において人手によって入力されたデータである．校正していないと思われる箇所が多く，文字誤りが多い．対訳文対は約 58,000 文対を含んでいる．現在産業技術総合研究所から入手することが可能である．ただし，研究目的に限る．そして，使用のための誓約書を産業技術総合研究所に提出する必要がある．

3.13 斎藤和英大辞典 [11] (AN)

分類は 1. 1938 年に故斎藤秀三郎氏が出版した辞書．見出し語 5 万，用例 12 万，総頁数 4640 頁，当時としては前例のない大和英辞典である．斎藤は「日本人の英語はある意味で日本化されなくてはならない」と，当時としてはユニークな見解を述べており，そのため非常に癖のある辞書となっている．使用されている日本語は，やや古めかしく，かつ差別用語が含まれる．英文も意訳（一部は超訳）が多く，他の辞書に見られない個性豊かな辞書である．賞賛する研究者も多い．

なお，斎藤秀三郎は昭和 4 年に 64 歳で亡くなっているため，この辞書の著作権は切れている可能性がある．ただし，斎藤氏の原稿は 「H 」の項までしか作成していないらしい．なお，対訳文対は用例ファイルとして本文と別になっているため，対訳文対の採取は容易である．

3.14 小倉書店 英語文型・文例辞典 [12] (AO)

分類は 7. 文単位の英日対訳文対が販売されている希少な例である．なお，原文はテキストではなく HTML 形式である．自然科学系の論文，報告書，仕様書あるいは書簡文などの構成に必要な表現例を項目別に編集したコーパスである．

3.15 英辞郎 用例コーパス [13] (AP)

分類は 7.「英辞郎」とは，プロの翻訳者・通訳者で構成されるグループ (EDP) が制作する英和辞書の名称である．英辞郎には，一般的な単語はもちろんのこと，スラング，イディオム，ビジネス用語，経済用語，法律用語，特許用語，コンピュータ用語，科学技術用語，医学用語，固有名詞 (組織名・企業名・人名・国名・映画名) などが含まれている．現在見出し数では，150 万を超えていて，日本最大の辞書になっている．ただし，句の対訳が多く，対訳文対は少ない．

3.16 研究社 新編英和活用大辞典 [14] (AQ)

分類は 1. 例文は比較的分かりやすい規則で収録されている．ただし，得られた対訳文対と公表されている用例数には，大きな差がある．

3.17 ランダムハウス英語辞典 [15] (AR)

分類は 2. このコーパスは独自のフォーマット形式を持っている．しかし，EPWING のフォーマットに変換するツールが Internet で公開されているため，対訳文対を採取することができる．また，英文は高品質で，英語の native でなければ思いつかない文があるが，日本文は，日本語の native にとって奇異な文がある．

3.18 ビジネス技術実用英語大辞典 [16] (AS)

分類は 1. 用例が別ファイルになっている．このコーパスの対訳文対は，他のコーパスと比較すると，理系向きの文章が多く，かつ高品質である．

3.19 コンピュータ用語辞典第 3 版 [17] (AT)

分類は 1. カタログには “例文 12,600 件（延べ）を収録” とあるが，これは英和，和英それぞれの例文の合計を表していると思われる．英和の例文と和英の例文は大部分が重複している．

3.20 佐良木コーパス (AU)

分類は 8. 日本大学の佐良木昌氏が個人的に収集した対訳文対である．

3.21 白井コーパス (AV,BD)

分類は 8. 元 NTT-AT の故白井諭氏が個人的に収集した対訳文対である．

3.22 斎藤健太郎コーパス:比較構文 (AW)

分類は 6. 元鳥取大学工学部知能情報工学科池原研究室の斎藤健太郎氏が，様々な本から比較構文のみを収集した対訳文対である．

3.23 澤田康子コーパス:因果関係構文 (AX)

分類は 6．元鳥取大学工学部知能情報工学科池原研究室の澤田康子氏が，様々な本から因果関係構文のみを収集した対訳文対である．

3.24 アルク 英語教師用データベース [18] (AY)

分類は 4．アルクが公開している英語教師用の対訳文対である．ただし，現在公開が中止されている．

3.25 研究社 総合ビジネス英語文例事典 [19] (AZ)

分類は 1．この辞書では例文が複数行で掲載されている，そのため，対訳文対の採取はかなり複雑になる．

3.26 新実用英語ハンドブック [20] (BA)

分類は 1．この辞書では例文が複数行で掲載されているため，対訳文対の採取はかなり複雑になる．

3.27 研究社 新和英大辞典 [21] (BB)

分類は 1．この辞典の例文は比較的分かりやすい規則で収録されている．公表されている例文数は約 5 万であるが，採取すると約 20 万文対が得られる．

3.28 三省堂 エクシード英和辞典 [22](BE)

分類は 8． 元 NTT-AT の故白井諭氏が個人的に収集した対訳文対である．辞典を人手により入力したコーパスである．

3.29 科学技術日英・英日コーパス辞典 [23](BF)

分類は 2．独自のフォーマット形式である．今回採取した辞書のなかでフォーマットの情報が最も少なかった．用例は多い．

3.30 日本語文型辞典 [24](BG)

分類は 8．日本語文型辞典は，外国人が日本語を勉強するために書かれた日本語の例文集である．この例文を鳥取大学の故池原悟氏が CREST[池原 (2000] の費用で英訳した．

3.31 旺文社マルチ辞書 辞ショック [25](BH)

分類は 1．

3.32 田中コーパス [26](BI)

分類は 6．元兵庫大学の故田中康仁氏が学生と作成した対訳文対である．日本文から英文を作成している．主に学生が翻訳したため，英文の品質にバラツキがある．対訳文対の量は，約 20 万文対ある．この対訳文対は，過去に Internet で一般に公開されていた．そして，誓約書を書くことで全対訳文対が入手可能であった．しかし，現在公開されていないようである．

3.33 読売新聞社説 (BJ)

分類は 8．人手によって読売新聞の社説と The Daily Yomiuri と比較して，文単位に編集された対訳文対である．日本大学の佐良木昌氏が個人的に収集したコーパスである．

3.34 アルク 英語表現辞典 [27](CA,CB,CC,CD,CE,CF,CG)

分類は 4．このコーパスは，7 つに分類される．

3.35 英文ビジネスレター実用フォーマットと例文集 [28] (CH)

分類は 3．高島康司著．数少ない CD-ROM 付の日英の対訳のある書籍である．簡単なスクリプトを作ることで，対訳文対が採取できる．ただし，得られる対訳文対は少ない．

3.36 英文 E メール文例集 [29](CI)

分類は 3．向井京子著．数少ない CD-ROM 付の日英の対訳のある書籍である．簡単なスクリプトを作ることで，対訳文対が採取できる．ただし，得られる対訳文対は少ない．

3.37 読売新聞記事 [30](CK,CL)

分類は 5．読売新聞の英字新聞として The Daily Yomiuri がある．これらは，別々に販売されているが，同じ内容の記事が載っている．しかし，両者の関係は，対訳文対どころか，記事単位の対訳にすらなっていない．しかし，このようなコーパスから文単位の対訳文対を自動的に採取する研究があり，これを利用して，約 20 万文対の対訳文対が作成されている [Utiyama (2003)]．そして，読売新聞の記事を購入することを前提に，入手可能である．ただし，読売新聞の CD-ROM は非常に高価である．1 年分の日本文記事がアカデミック価格で 12 万円，英文記事が 11 万円である．また記事対応になっていないテキストから自動的に対訳文対を採取しているため，他のコーパスと比較すると対訳文対になっていない文対が多い（5.1 節参照）．つまり，日本文から見ると，英文の誤りが多い．

3.38 ATR バイリンガル旅行会話データベース [31](CM,CN)

分類は 7． ATR が音声対話システムのために収集した音声と対訳のコーパスである．対話総数 618，発話総数 16,107 文対で構成されている．

3.39 NHK やさしいビジネス英語実用フレーズ辞典 [32](CO)

分類は 3．この辞典は NHK ラジオ講座「やさしいビジネス英語」（杉田 敏）の Vocaburary Building の内容をまとめた本で，CD-ROM 中に対訳文対を含んでいる．

3.40 自然科学系和英大辞典 [33](CQ)

分類は 1．

3.41 ジーニアス英和・和英辞典 [34](CR)

分類は 1．このコーパスでは例文の中の見出しの単語や連語は “〜” で略されている．そのため，対訳文対を採取しても単語を対応づけることが困難であるため，誤りのない対訳文対を採取することは困難である．本研究では，“〜” を見出しの単語に置き換えて，対訳文対を作成した．

3.42 最新ビジネス英文手紙辞典 [35](CS)

分類は 3．このコーパスは，英文の手紙の例文を収集している．

3.43 機械を説明する英語 [36](CT)

分類は 7．

4　得られた対訳文対の量

4.1　得られた対訳データベースの文対の数

得られた対訳文対の数を表 11 中の「採取文数」に示す．この数字は，文章と推定して採取した文対の数であり，クリーニングが完全でないため，誤った対訳文対が含まれている．様々なコーパスから対訳文対を採取して，総計として 1,099,093 対訳文対を採取した．これらのコーパスでは，約 70% が単文と認定できる文であった．約 20% は重文・複文と認定できる文であった．そして残りの約 10% は．複雑な重文・複文で文長が長い文であった．また大多数は，テキスト文であるが，一部には，対話文があった．対話文の多くは旅行会話である．なお，本論文では，この対訳文対を対訳データベースと呼ぶ．

4.2　得られた単文データベースの文対の数

採取した対訳データベースから，日本文において単文の対訳文対を採取する．一般的には，単文の定義は「述語が一つだけから成る文」であるが，この条件では定義できない文が多い．本研究では，単文の条件を以下のように定義する [西山 (2005)]．

1. 文末以外に動詞が一つもなく，文末が動詞で終わる文
   * 彼は毎日自転車に乗る。
2. 文中に動詞がなく，文末が複合動詞で終わる文
   * ドイツは新しい歴史への一歩を踏みだした。
3. 文中に動詞，複合動詞，形容詞が一つもなく，文末以外に形容詞が一つもなく文末が形容詞で終わる文
   * この林檎はややすっぱい。
4. 文中に動詞，複合動詞，形容詞が一つもなく，文末以外に形容動詞がが一つもなく文末が形容動詞で終わる文
   * 企業の経営戦略は大切だ。
5. 文中に動詞，複合動詞，形容詞，形容動詞が一つもなく，文末が “名詞＋付属語” で終わる文
   * あの人こそ真の英雄だ。
6. 疑問文，命令文，会話文は対象外
   * 疑問文：この本は何について書いてあるか。 * 命令文：そこに私のテントを張れ。 * 会話文：昨日どこかへ行ったかい。

また，日本語で分類を行っているため，英文は複文になっている対訳文対がある．本論文では，この対訳文対を単文データベースと呼ぶ．得られた単文データベースの文対の数を表 11 の「単文」に示す．また，抽出した単文の例を表 1 に示す．

単文データベースは，できるだけ簡潔で問題のない単文のみを選択したため，対訳データベース（4.1 節）の多くの単文が，利用されなかった．最終的には，単文データベースは総計で 182,113 対訳文対となった．ただし，単文データベースは，全ての文対を人手でクリーニングしていない．

表 1　単文データベースの例

日本文 1	猫が縁側をのそのそ歩いている。
英文 1	The cat is walking across the veranda .
日本文 2	新宿のネオンサインが消えた。
英文 2	The neon sign lights of Shinjuku were turned off .
日本文 3	銃声が一発聞こえた。
英文 3	A loud report of a gun was heard .

4.3　得られた重文・複文データベースの文対の数

採取した対訳文対から，人手によって文を解析して，重文および複文を採取した．ただし，分類は日本文で行っている．そのため英文は単文になっている対訳文対がある．採取した重文・複文の文対の数を表 11 の「重文・複文」に示す．総計として 158,633 対訳文対を採取した．なお，本論文では，この対訳文対を重文・複文データベースと呼ぶ．

4.4　重文・複文の文種別

採取した重文・複文データベースを，以下に示す文種別 1 から 5 に従って人手で分類した．分類は日本文で行っているため，英文が単文である対訳文対もある．分類して得られた文対の数を，表 11 中の文種別 1 から 5 に示す．ただし，埋め込みについて修飾要素を持たない用言（連体形）は埋め込み文としない．

1. 文種別 1 （重文）
   文接続を一つ持つ文である．いわゆる重文である．例文を表 2 に示す．

表 2　文種別 1 の例

日本文 1	窓を開けると冷たい風が入ってくる。
英文 1	When you open the window a cold wind blows in.
日本文 2	彼女に会ったおかげで一日が楽しかった。
英文 2	Seeing her made my day.
日本文 3	もし行きたいのなら行きなさい。
英文 3	You can go if you choose to go.

2. 文種別 2 （重文）
   文接続を二つ持つ文である．いわゆる重文である．例文を表 3 に示す．

表 3　文種別 2 の例

日本文 1	昔の友達に久しぶりに会って、夜を更かして語り合った。
英文 1	I met an old friend for the first time in a long time, and we chatted late into the night.
日本文 2	私はカギをなくしてしまったので、妻が帰るまで、待たなければならなかった。
英文 2	I lost my key, so that I had to wait till my wife returned.
日本文 3	その肉はよく煮ないと、かたくて食べられない。
英文 3	If you do not cook this meat well, it will be too tough to eat.

3. 文種別 3 （複文）
   埋め込み文を一つ含む文である．いわゆる複文である．例文を表 4 に示す．

表 4　文種別 3 の例

日本文 1	それを知らぬ者は誰もいない。
英文 1	There is no one but knows that.
日本文 2	誰も完全に幸福な者はいない。
英文 2	None are completely happy.
日本文 3	船が水平線以下に隠れるまで見送った。
英文 3	I followed the ship with my eyes till she disappeared below the horizon.

4. 文種別 4（複文）

埋め込み文を二つ含む文である．いわゆる複文である．例文を表 5 に示す．

表 5　文種別 4 の例

日本文 1	通りの方へ向かっている窓と中庭に向かっている窓がある。
英文 1	Some windows look out on the street, the others look out into the yard.
日本文 2	山を汚す者に山を楽しむ資格はない。
英文 2	People who leave trash in the mountains are not qualified to enjoy them.
日本文 3	彼がこつこつと金を貯めているのは、外国へ旅行するためです。
英文 3	He is diligently saving money in order to travel overseas.

5. 文種別 5 （重複文）

文接続を一つと埋め込み文を一つ含む文である．いわゆる重複文である．例文を表 6 に示す．

表 6　文種別 5 の例

日本文 1	ドアの開く音がかすかに廊下に響いた。
英文 1	The sound of the door opening echoed faintly down the corridor.
日本文 2	昔のことを思い出すと重苦しい悲しみが彼女の心をおおった。
英文 2	A leaden grief swept over her at the thought of her past.
日本文 3	彼は家を建てるために節約してお金を貯めている。
英文 3	In order to build a house, he is economizing and saving money.

5　考察

5.1　採取した対訳文対の誤り調査

採取した対訳文対の精度を調査するために，人手による誤り調査を行った．単文データベースからランダムに 100 文対抽出した．この 100 文対を調査したところ， 4 文対に誤りが検出された．誤りが検出された文を表 7 に示す．

出典を調査したところ，日本文 1 は講談社和英辞典で，日本文 2 はランダムハウス英語辞典で，日本文 3 と日本文 4 は，読売新聞（文対応データ）であった．講談社和英辞典は，入力誤りが多いコーパスである．ランダムハウス英語辞典は，特殊フォーマットであるため，採取に誤りが生じたと考えている．また，読売新聞（文対応データ）は，記事対応の日英対訳文から，プログラムで，自動的に文対応の対訳文対を作成している．そのため誤りが多いと考えている．

なお，重文・複文データベースは，全文を人手により検査して修正を行っている．そのため，重文・複文データベースを単文データベースと同様に調査したが，誤りは見つからなかった．

表 7　単文データベースにおいて誤りが発見された文対

日本文 1	どんな もの か 頓 と 見当 も つか ない 。
英文 1	I have no ideaof what it is like . （“idea of” の誤り.）
日本文 2	彼 から 事業 全体 を 残ら ず 買い取っ た 。
英文 2	We bought the whole business from him,lock,stock , and barrel . （“,lock,stock , and barrel” が不要）
日本文 3	国 の 工業 用 エタノール 買い取り 価格 の 半額 という 安 さ だ
英文 3	The expected price of the produced ethanol will be only half of the 100,000 yen that the government pays to purchase a kiloliter of industrial ethanol , the sources said . （英文が明確に誤っている.）

5.2　対訳文対の品質の問題

採取した対訳文対には，対訳文対として不適切と思われる文対が存在する．意味は解るが英文の品質に問題があると思われる．例文対を表 8 に示す．なお，欠陥英和辞典の研究 [副島 (1989)] において，辞書の例文には，英文として適切ではない例文が存在することが報告されている．

表 8　対訳文対の品質が問題になる例

日本文 1	嫌悪の叫びをあげた。
英文 1	She cried out in revulsion. (日本語に “彼女は” がない.）
日本文 2	顧客サービス部をお願いします。
英文 2	Give me Customer Service. （“Customer Service” は，電話対応における “顧客サービス部” であり，対面の場合は正しいのか？）
日本文 3	皇太子ご夫妻は九三年六月に結婚された。
英文 3	The crown prince and the princess married in June 1993 . （日本語において一九九三年とすべきである．）

本研究では，可能な限り多くの電子媒体から，可能な限り品質の高い対訳文対の採取を試みた．しかし，精度の高い対訳文対を大量に得ることは困難であり，問題のある対訳文対と誤りのある対訳文対の混入は避けられなかった（5.1 節）．著者の意見として，翻訳品質の高い対訳文対を 100 万文対，収集することは，かなり困難であると考えている．

5.3　辞書間の類似文の存在

採取された例文を調査すると，異なる辞書において，似た文章が多く掲載されている，例えば，複数の辞書において，英文で “Wine is made from grapes.” を検索した日本文を表 9 に示す．このような例が多くの辞書において見られる．

表 9　“Wine is made from grapes.” の対訳文

ワインはぶどうから作られる	ワインはぶどうでつくる	ブドウ酒はブドウからつくられる
ワインはブドウから作る	ぶどう酒は葡萄より作られる	ワインはブドウで作る

5.4　過去の辞書の例文における著作権の問題

過去に辞書に集録された例文が問題になった例として，以下の事例が報告されている [副島 (1990)]．1967 年に発行された研究社・新英和中辞典（初版）（岩崎民平・小稲義男）は，開拓社の新英英大辞典 (ISED) とオックスフォード辞書現代英英辞典 (OALD) の例文を，大量に引用した．そのため，ISED と OALD の著者である A.S.Hornby と ISED の発行元である開拓社から抗議をうけて，翌 1968 年，例文を変更して，研究社・新英和中辞典第 2 版が出版された．

6　コーパス等の外部提供等に関する著作権法等との抵触について [藤波 (2012)]

6.1　はじめに　ー著作権問題ー

デジタル化・ネットワーク化の急速な進展に伴う著作物の利用形態には，既設の著作権法権利制限規定により可能である行為と実質的に同様の行為も多いのですが，権利制限規定が個別具体の事例に沿って定めていることから，たとえ権利者の利益を不当に害しないものであっても形式的には違法となってしまうとの指摘（「デジタル・ネット時代における知財制度の在り方について」検討経過報告：平成 20 年 5 月 29 日，同．報告：平成 20 年 11 月 27 日，知的財産戦略本部／デジタル・ネット時代における知財制度専門調査会）がなされており，これに対応する著作権法改正が行なわれ平成 22 年 1 月 1 日に施行されました．

この改正著作権法では，研究開発における情報利用の円滑化を図る目的での権利制限規定も新設され，例えば，画像・音声・言語・ウェブ解析技術等の研究開発過程での著作物等の利用について著作権法上の問題が生じるなどの指摘に対して，これを適法化する立法的解決が図られています．

しかし，新設された情報解析のための複製等の権利制限規定（著作権法：第四十七条の七　以後”著 47 条の 7”と略します．）の文言は，著作物その他の記録または創作した二次的著作物の記録を含む翻案への言及に留まり，作成したコーパスの外部提供等やコーパスが依拠した記録の外部提供等についての明文規定は省かれています．このため，著 47 条の 7 等に依拠して作成したコーパスの外部提供等や利用について，主に技術系研究者から様々な意見—-「公開不可で第三者がダウンロードして評価することができない」，「どのような利用方法まで許容されるのか不明であり，確定判決（最高裁判決？）を得るまで他者は利用できない」等々—-が出され，情報解析のための複製等の権利制限規定（著 47 条の 7）の立法趣旨は勿論，著作物利用に係る判例などもが没却されているようにもみえます．

本稿では，著 47 条の 7 の解釈や適用範囲を画する立法事実，著作物性がなく著作権法の保護を受けない表現の利用行為，行為の外形上は著作権法の保護対象となる利用行為に当たるものの著作権法が保護すべきものとして本来想定しているような利用行為とは利用形態が異なる利用行為，違法の疑義がある利用行為についての適正手続き（deu process of law）について，文化審議会資料や関連判例などに依拠し，作成したコーパスや依拠した記録の外部提供等などに係る著作権法等との抵触の存否と適法性を確保する要件について述べます．

なお，著 47 条の 7 の解釈等にかかる考察につき「私見が述べられているにすぎない…」との意見もありますが，法令の公権的解釈権は裁判所が有し，その解釈は具体的争訟性事件性のある法律上の争訟について行なわれる（裁判所法 3 条）日本の司法制度下では，裁判「所」以外が示す法令の解釈や意見は全て「私見」であり，本稿も同様です．

6.2　著作権法：（情報解析のための複製等）第四十七条の七 (著 47 条の 7)

著作権法：（情報解析のための複製等）第四十七条の七を以下に示します．

著作物は，電子計算機による情報解析（多数の著作物その他の大量の情報から，当該情報を構成する言語，音，映像その他の要素に係る情報を抽出し，比較，分類その他の統計的な解析を行うことをいう．）を行うことを目的とする場合には，必要と認められる限度において，記録媒体への記録又は翻案（これにより創作した二次的著作物の記録を含む．）を行うことができる．ただし，情報解析を行う者の用に供するために作成されたデータベースの著作物については，この限りでない．

6.3　情報解析技術の研究開発に着目した権利制限およびその射程範囲

著 47 条の 7 に係る立法者の意思を以下の節に示します（文化審議会著作権分科会報告書 平成 21 年 1 月），本条項の適用・解釈は，「これらの文化的所産の公正な利用に留意しつつ，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」（著 1 条）ことを目的として，以下（1,2 節）に沿った適用・解釈が行なわれることになります．

1. 情報解析技術の研究開発に着目した権利制限の根拠
   （文化審議会著作権分科会．報告書 88 頁）
   高度情報化社会の下で，取り扱われる情報量が爆発的に増大する中，利用者が必要とする情報・知識を抽出し，高度な知的処理を実現する情報解析技術は，デジタル・ネットワーク社会の基盤となるものであり，そのための研究開発も社会的に意義を有する．
   情報解析分野の研究開発は，著作物の表現そのものを利用するものではなく，その情報・アイディアの抽出を行うに過ぎないが，その過程で中間的に利用行為に当たる行為を伴うものであり，著作物利用の実質を備えないとの側面もある
2. 権利制限を行う場合の要件
   （文化審議会著作権分科会．報告書 88-91 頁）
   （a）営利・非営利の別
   非営利のものに限定する必要はないと考えられる．その場合に著作権者等の利益が害されるおそれがあるとするならば，次の要件設定ー著作権者等の利益への影響（6.3 節 2b）ーより対応すべきである．
   （b）著作権者等の利益への影響
   契約によって入手可能なデータベース等の場合には権利制限を認める必要はない．このような意見に照らせば，既存のビジネスの中で研究開発に必要なデータベース等が有償で提供されているような場合，その他，著作物の性質や利用態様等に応じて著作権者等の利益を害すると考えられるような場合には，権利制限の対象外とすることが適当と考えられる．
   （c）研究開発の過程で作成された複製物の外部提供等
   権利制限が情報抽出のための過程で中間的に行われる複製であることに着目したものであるとの側面からは，基本的に，当該複製物を外部に提供することはその趣旨に反することになるため，当該複製物を研究に参加しない者に提供する行為については権利制限の対象外とすべきと考えられる．なお，研究過程で作成された複製物の外部提供の取扱いと関連して，研究開発を行う者にそのためのデータベース等を提供するような事業があった場合にこれが権利制限の対象となるかどうかについては，このようなデータベース等の作成自体が研究開発目的の者によって行われているかどうかで判断すべきとの指摘がある．

6.4　著作物性のない著作物（非保護著作物）等の利用

表現物の利用については，「既存の著作物に依拠して創作された著作物が，思想，感情若しくはアイディア，事実若しくは事件などの，表現それ自体でない部分又は表現上の創作性がない部分において，既存の著作物と同一性を有するにすぎない場合には，翻案には当たらない」（最一小判平成 13.6.28 江刺追分事件）ことから，思想，感情，アイディアの表現（例：大阪高判平成 6.2.25 脳波数理解析論文事件，東京高判平成 12.3.29 エスニシティ論文事件）はもちろん，著作物性が認め難い，短い表現（例：東京地判平成 13.5.30 チャイルドシート事件，知財高判平成 17.10.6 ヨミウリオンライン事件），ありふれた表現（例：東京地判平成 7.12.18 ラストメッセージ in 最終号事件），選択幅が狭い表現（例：東京地判平成 13.1.23 多摩地図事件，知財高判平成 20.7.17 ライブドア裁判傍聴記事件，東京地判平成 10.11.30 版画写真事件），機能表現（例：東京地判平成 15.1.31 電車線設計プログラム事件，大阪地判平成 4.4.30 丸棒矯正機設計図事件）などは，原則として著作権法では保護されず，利用することができます．また，「著作物の創作的特長を感得できないときは，複製等の著作物の利用には該当しない」とされ，著作物を利用している実態がないときは著作権は動かない（違法ではない）とされています（雪月花事件：東京地裁平成 11 年 10 月 27 日判決，はたらくじどうしゃ事件：東京地裁平成 13 年 7 月 25 日判決）．

更に「複製とは，既存の著作物に依拠し，その内容及び形式を覚知させるに足りるものを再製すること」（最一小判昭和 53.9.7 ワン・レイニー・ナイト・イン・トウキョー事件）とされ，「言語の著作物の翻案とは，既存の著作物に依拠し，かつ，その表現上の本質的な特徴の同一性を維持しつつ，具体的表現に修正，増減，変更等を加えて，新たに思想又は感情を創作的に表現することにより，これに接する者が既存の著作物の表現上の本質的な特徴を直接感得することのできる別の著作物を創作する行為」と定められている（最一小判平成 13.6.28 江刺追分事件）ことから，これらの要件のどの 1 つを欠いても，その利用行為は翻案にも複製権侵害にもなりません．

6.5　論点と対応

情報解析の研究開発における既存著作物等の利用目的は，著作物の思想，表現そのものを感じ取るのではなく，その中から研究開発に必要な部分を探し当てること，アイディアや背景情報等を抽出することなどであって，人間が行ったとするならば視聴行為として著作権が及ばない行為です．しかし，同様の行為をコンピュータ等に実行させるときは，中間的に既存著作物等を蓄積する必要があるために，物理的には，複製行為や翻案行為が行われることになります．著 47 条の 7 はこれらの行為を，権利制限規定を新設することで適法化して立法的解決を図ったものです．

著 47 条の 7 は「著作物は，電子計算機による情報解析（多数の著作物その他の大量の情報から，当該情報を構成する言語，音，影像その他の要素に係る情報を抽出し，比較，分類その他の統計的な解析を行うことをいう．…）を行うことを目的とする場合には，必要と認められる限度において，記録媒体への記録又は翻案（これにより創作した二次的著作物の記録を含む．）を行うことができる．」と規定し，情報解析目的での記録媒体への記録又は翻案を適法としています．但し，「必要と認められる限度において」要件を設けており，「必要と認められる限度」とは何かが論点となります．なお，ただし書きでは「情報解析を行う者の用に供するために作成されたデータベースの著作物については，この限りでない．」と定め，立法時に付された権利制限の要件（6.3 節 2b 参照：契約によって入手可能なデータベース等の場合には権利制限を認める必要はない）を明記しています．

著 47 条の 7 は本条に依拠して作成したコーパス，研究開発の過程で作成された複製物および依拠した記録の外部提供についての明文規定はなく，これらの外部提供（の可否や要件）が論点となります．また，著作権，著作者および実演家の人格権，著作隣接権等に対する侵害は親告罪とされており，その侵害について刑事責任を追及するかどうかは被害者である権利者の判断に委ねられ，犯人を知った日から 6 か月を経過したときは告訴することができませんし（刑訴法 235 条：告訴期間），被害者の被害感情や被害の重み，訴追意思は，公訴提起の判断において重視され，一般に，被害者の意思と全く無関係に訴追が行われることはありません（同．248 条：起訴便宜主義）．従って，当該利用方法が予期しなかった権利侵害の疑義が生じたときは，権利者への対応や権利侵害（民法 709 条：不法行為）を防止する措置の有無や内容により侵害認定が異なることから，対応や防止措置が論点になります．

以下，著 47 条の 7 に係る，次の論点について検討します．

論点 1　必要と認められる限度

「研究開発における情報利用の円滑化」との課題に対応して新設された著 47 条の 7 を，著作権法が本来想定している保護範囲と外形上はその利用行為に当たるものの利用の実質を備えない行為との調整である（文化審議会著作権分科会報告書 86 頁）と解すると，「必要と認められる限度」は，権利の例外を設ける際の一般的要件の充足が要件となります．著作権関係条約における権利例外を設ける際の基準は，一般的には表 10 に示す「スリーステップテスト」が基準になっています．

表 10　スリーステップテストの 3 要素

1	特別の限定された場合であること，	2	通常の利用を妨げないこと，	3	権利者の利益を不当に害しないこと

著 47 条の 7 に依拠する記録媒体への記録又は翻案は，この 3 要件を充足する限度で行為することが求められています．

なお，営利・非営利の要件については，本件権利制限の根拠を情報解析技術に関する研究開発の社会的意義等に求める考え方に照らせば，非営利のものに限定する必要はありません．ただし，著作物の性質や利用態様等に応じて著作権者等の利益を害すると考えられるような場合には，権利制限の対象外とすることが適当と考えられます（同．88-89 頁）．

論点 2　研究開発の過程で作成された複製物の外部提供

本件権利制限が情報抽出のための過程で中間的に行われる複製であることに着目した側面とその趣旨から，当該複製物を研究に参加しない者に提供する行為は権利制限の対象外とすべきと考えられます．なお，研究開発を行う者にそのためのデータベース等を提供するような事業があった場合に，これが権利制限の対象となるかについては，このようなデータベース等の作成自体が研究開発目的の者によって行われているかどうかで判断すべきである（同．89 頁）ことから，研究開発目的の者による研究開発の過程で作成された複製物の外部提供は適法と解され得ます．ただし，この場合であっても，権利例外を設ける際の基準である「スリーステップテスト」の基準を充足した外部提供方法であることを要します．

なお，情報解析のために複製・翻案した複製物や二次的著作物の保存を禁止する著作権法条項はありませんので，複製物や二次的著作物の保存は適法行為と解されます

また，元の複製者が，作成された著作物の複製物や二次的著作物を送信可能な状態にすることを禁止する条項もありませんので，送信や送信可能化する行為も適法と解されます．しかし，複製物・二次的著作物を送信可能な状態におくことは，受信者の目的が法で定められたものであるかの確認ができないことから，当該物を送信可能な状態にした者の不法行為責任（民法 709 条）が生じることがありますので，受信者の目的確認および相応の使用契約に合意した者だけが利用できるなどの措置を採ることが必要になります．

論点 3　コーパスの外部提供

コーパスは，情報抽出過程で中間的に存在する記録から抽出された表現などで構成されています．コーパスはデータベースの著作物であり，通常は，コーパス作成者がデータベースの著作権者であり，コーパスの部分を構成する表現が権利処理された保護著作物または非保護著作物であるときは，当該コーパスを外部提供することができます．

コーパスが著 47 条の 7 に依拠して作成されたときは，その複製物および（作成されたコーパスが新しい著作物か二次的著作物かの論点はあるが）二次的著作物の目的外使用が問題になります．即ち，「第四十七条の七に定める目的以外の目的のために」著作権制限規定の適用を受けて作成された著作物の複製物を利用した場合（著 49 条第 1 項第 5 号）あるいは，同条の規定適用を受けて作成された二次的著作物を利

用した場合（著 49 条第 2 項第 6 号）は違法（複製権侵害）とされます．ただし，著 49 条の文言が「…を利用した者」となっていることに留意した対応が必要になります．

以上により，「第四十七条の七に定める目的」でのコーパスの外部提供は適法行為と解されます（著 49 条第 1 項第 5 号，同. 第 2 項第 6 号の反対解釈）．また，外部提供行為は非営利のものに限定する必要はないと考えられます（文化審議会著作権分科会報告書 88 頁）．

論点 4　コーパスが依拠した原記録の外部提供

コーパスの的確な利用には，コーパス作成に係る原表現記録から当該コーパスの利用領域や利用可能深度を検討する必要があります．故に，言語解析の研究開発過程では，単語や文のつながりなどの用例をウェブ上で検索・表示可能にすることが行なわれて，機械翻訳等に関する研究開発，辞書・文法書の編纂，言語研究等にコーパスが用いられています（文化審議会著作権分科会報告書 85 – 86 頁）．著 47 条の 7 は，著作権法が本来想定している保護範囲と外形上はその利用行為に当たるものの利用の実質を備えない行為との調整である（同．85-86 頁）ことから，著 47 条の 7 に依拠する複製物をウェブ上で検索・表示可能にするなどの行為は適法行為と解されます．但し，本件複製は「必要と認められる限度」で認められることから，検索・表示可能にする記録は「スリーステップテスト」の基準を充足する限度内であることが必要です．

論点 5　外部提供に係る他人の権利や法律上保護される権利の侵害防止

コーパスなどの外部提供ではコーパス作成者も予期しなかった権利侵害の疑義が生じて，正当な権利者や法律上保護される利益を有する者から侵害防止の措置等が求められることがあります．このため，著作権者から保護著作物の削除要求（プロバイバー責任制限法類推）があった場合の削除手続き（オプトアウト：Opt-Out．個人情報保護法 23 条類推）や外部提供に係る権利侵害の申し出手続きと外部提供者が採る権利侵害防止措置などを予め定めて，それらを利害関係人が知り得る状況を作出しておく必要があります．また，著作権法が親告罪であることに鑑み，特にコーパスが依拠した原記録の外部提供に際しては，提供の趣旨，提供情報の非保護著作物化，適法な提供先であることの確認方法など，提供時の適正手続きの制定や実行を可視化して，利害関係人の事前の理解を得ることも必要になります．

6.6　まとめ　ー著作権問題ー

著 47 条の 7 等に係るコーパス等の外部提供等に関する著作権法等との抵触に係わる論点と結論を以下に示します．

論点 1　記録媒体への記録又は翻案における「必要と認められる限度」

「スリーステップテスト」（表 10）基準の 3 要件を充足する限度に限られます．

論点 2　研究開発の過程で作成された複製物の外部提供

研究開発目的の者による研究開発の過程で作成された複製物に限り，研究に参加した者および研究開発を行なう者に限定して，提供することができます．ただし，その提供態様が「スリーステップテスト」の 3 要件を充足しないときは，この限りではありません．

論点 3　コーパスの外部提供

著 47 条の 7 が定める目的に限定して，外部提供できます．また，提供は非営利のものに限定されません．

論点 4　コーパスが依拠した原記録の外部提供

ウェブ上で検索・表示可能にするなどの態様で，外部提供できます．ただし，検索・表示可能にする記録は「スリーステップテスト」の基準を充足する限度内であることが必要であり，それを担保する相応のシステム対応などが必要になります．

論点 5　外部提供に係る他人の権利や法律上保護される権利の侵害防止

権利者から侵害防止の措置等が求められた場合の対応措置（情報削除等の申出・実行・確認等の手続きなど）を予め定めて公開するなど，侵害防止のための適正手続きが必要になります．

6.7　おわりに　ー著作権問題ー

言語解析などの情報解析技術の研究開発で求められていた著作権の権利制限規定等の措置は，平成 22 年 1 月 1 日施行の改正著作権法で，ほぼ終えたものと解されており，新設された権利制限規定等をどのように解し用いて研究開発を推進し成果を社会に供するかは技術者側の対応に委ねられています．もちろん，これらの規定に明文の規定がなく個別具体的な事例に即し判例などで定まる論点もありますが，ある限度で結論が定まる論点がほとんどであり，その限度内で研究開発を進めることが求められています．

なお，コーパス等の外部提供等については，幇助，間接侵害，プロバイダ責任制限法との関係など余の論点も多々ありますが，本稿では割愛します．また，本稿は法律専門語で論じるべき問題を通常の言葉で記述していることから，用いる語彙や意味，論旨展開が法学で用いるモノと異なる箇所があり，内容も法学的厳密性を必ずしも具備していないことに留意してください．

7　まとめ

日英対訳辞書は，翻訳の研究において必要不可欠のものである．しかし，日本文と英文が文単位に対応していて，量が多く，一般の人が入手可能な対訳文対は，最近まで存在していなかったと言える．

本研究では，様々な電子媒体から，日本文と英文が文単位に対応する対訳文対を採取した．電子媒体として，CD-ROM・Internet・電子辞書などを利用した．これらの結果，対訳文対として 1,099,093 万文対を採取した（対訳データベース）．そして，得られた対訳文対から，日本文において単文の対訳文対を 182,113 文対，採取した（単文データベース）．また日本文において重文・複文の対訳文対を 158,633 文対，採取した（重文・複文データベース）．また，重文・複文の対訳文対を採取した重文・複文データベースを，文種別 1 から 5 に従って人手で分類した．

ただし，単文データベースは，コストの問題から全てを人手でクリーニングできていない．そのため 100 文対を調査したところ 4 文対の誤りがあった．一方重文・複文データベースは，全文を人手により検査して修正を行っている．そのため誤りが発見できなかった．しかし，日本文と英文を比較すると，翻訳精度の高い文対になっていない場合がある．これらを考慮すると，翻訳品質の高い対訳文対を 100 万文対，収集することは，かなり困難であると考えている．

なお，統計翻訳の目的のために，様々な電子媒体から対訳文対を採取することは，著作権法：(情報解析のための複製等）第四十七条の七から，問題ないと考えている．そして，採取した対訳文対を特定のグループ内で利用することも，許される行為と考えている．ただし，“必要と認められる限度” を，常識的に判断する必要がある．

謝辞

ここで紹介したコーパスは，長年翻訳に携わってきた多くの方々の努力を，著者が覚え書きとしてまとめたものです．日本大学の佐良木昌氏には，個人的に収集して頂いた対訳文対を利用させて頂きました．電子辞書からの対訳文対の採取には，当時鳥取大学工学部知能情報工学科の片山慶一郎氏（既卒）の助力を得ました．重文・複文の分類は，NTT-AT の，木村淳子氏，小見佳恵氏，阿部さつき氏，竹内（村本）奈央氏，小船園望氏が中心になって行いました．鳥取大学の徳久雅人氏および山梨英和大学（岐阜大学）の池田尚志氏には，この辞書の作成にあたり，様々な協力を得ました．以上の方々に感謝いたします．最後に，元兵庫大学の田中康仁氏と元鳥取大学（NTT) の池原悟氏と元 NTT-AT（NTT,ATR) の白井諭氏は鬼籍に入られました．ご冥福をお祈りいたします．

参考文献

[1] 池原悟, 白井諭, 小倉健太郎, “言語表現体系の違いに着目した日英機械翻訳試験項目の構成”, 人工知能学会論文, Vol.9, No.5, pp.569-579, 1994.

[2] 桑畑 和佳子, 橋本 三奈子, 村田 賢一, “計算機用日本語辞書の開発”, 情報処理学会研究報告, 人文科学とコンピュータ研究会報告 93(42), 27-34, 1993.

[3] “学研 ニューアンカー英和・和英辞典,” B298C49I1, 2000.

[4] “外国人のための基本語用例辞典 （第 3 版）” 文化庁国語課約 4,500 語, 大蔵省印刷局, ISBN 4-17-151302-2, 2000.

[5] “英語表現辞典”, 三省堂編修所, ISBN 4-385-11012-3, 1997.

[6] “英文ビジネスレター文例大辞典 Ver.2[CD-ROM]”, 日本経済新聞社出版局, ISBN 453245509X. 1997.

[7] “外国人のための日本語例文・問題シリーズ”, 荒竹出版.

[8] “自然発話音声・言語データベース（日英対訳)”, http://www.atr-p.com/sdb.html

[9] 黒橋禎夫, 白井清昭, ”SENSEVAL-2 日本語タスク”, 信学技報, NLC 101(351), 1-8, 2001.

[10] “講談社和英辞典”, 講談社, ISBN-13: 978-4061210530, 1982.

[11] 斎藤秀三郎, “斎藤和英大辞典”, 日外アソシエーツ辞書編集部編, EPWING , ISBN4-8169-8078-4, 1999.

[12] “英語文型・文例辞典”, 小倉書店, http://www.ogurashoten.co.jp/kyozai3.html, 1997.

[13] “英辞郎”, アルク, http://shop.alc.co.jp/cnt/eijiro/.

[14] “新編英和活用大辞典” 研究社, ISBN4-7674-3573-0, 1995.

[15] “ランダムハウス英語辞典 第二版 CD-ROM 版”, 小学館, http://ebook.shogakukan.co.jp/scatalog/randam/top/top.htm, 2002.

[16] 海野文男, 海野和子, “ビジネス技術実用英語大辞典”, 日外アソシエーツ, ISBN4-8169-8127-6, T4937695181270, 2000.

[17] “CD-コンピュータ用語辞典 第 3 版英和・和英/用例・文例パッケージ”, コンピュータ用語辞典編集委員会〔編〕, 日外アソシエーツ, ISBN4-8169-8126-8, T4937695181263, 2000.

[18] “英語教師用データベース”, アルク, http://home.alc.co.jp/db/owa/engt_structure?stg=4

[19] “研究社ビジネス英語スーパーパック”, 研究社出版, ISBN4-7674-3590-0, 1998.

[20] “新実用英語ハンドブック”, 大修館書店, ISBN4-469-74233-3, 1995.

[21] “新和英大辞典”, 研究社, ISBN 4-7674-7200-8, 2003.

[22] “エクシード英和辞典”, 三省堂編修所, ISBN 4-385-10650-9, 1998.

[23] “科学技術日英・英日コーパス辞典”, 丸善, ISBN4-621-04991-7, 2002.

[24] “日本語文型辞典”, くろしお出版, ISBN-10: 4874241549, ISBN-13: 978-4874241547, 1998.

[25] “旺文社版マルチ辞書 W 辞ショック”, 株式会社アスク, AWR1-00430, 1997.

[26] 田中康仁 (兵庫大学), “日英・パラレルコーパスの作成”, 言語処理学会 第 8 回 年次大会, B4-2, pp.499-502, 2003.

[27] “英語表現辞典”, アルク, http://www.alc.co.jp/eng/kaiwa/hyogen/index.html

[28] 高島康司, “英文ビジネスレター実用フォーマットと例文集”, ベレ出版, ISBN4-939076-25-3, 2000.

[29] 向井京子, “英文 E メール文例集”, 池田書店, ISBN4-262-16896-4, 2002.

[30] 内山将夫, 井佐原均, “日英新聞記事の対応付けと精度評価”, 第 151 回 自然言語処理研究会第 68 回 情報学基礎研究会, pp.15-22, 2002.

[31] 竹沢寿幸, 白井諭, 大山芳史, “バイリンガル旅行会話コーパスに見られる話し言葉の特徴分析”, 自然言語処理研究会報告, pp.137-144, 2001.

[32] 杉田敏 編, “NHK やさしいビジネス英語実用フレーズ辞典” NHK 出版, ISBN978-4-14-034102-5, 2003.

[33] “自然科学系和英大辞典” 小倉書店, ISBN-13: 978-4902764000, 1997.

[34] “ジーニアス英和・和英辞典 CD-ROM 版” , 大修館書店, ISBN4-469-79057-5, 2001.

[35] フランシス・J・クディラ (著), 朝日出版社, “最新ビジネス英文手紙辞典 CD-ROM 新訂版”, ISBN-13: 978-4255002927, 2004.

[36] “機械を説明する英語”, アスク, ASIN: B00008HV7V, 1999.

[37] “リーダーズ英和辞典”, 研究社, ISBN4-7674-3563-3, 1999.

[38] 木原研三, 小西友七, 他, “新グローバル＆ニューセンチュリー英和・和英辞典”, ISBN4-385-61400-8, JAN コード：T4938641614002, 1994.

[39] “ＣＤ- 科学技術４５万語対訳辞典”, 日外アソシエーツ, ISBN4-8169-8128-4, 2001.

[40] “新英和・和英中辞典”, 研究社, ASIN: B0009EWFA0, 2005.

[41] EPWING, http://www.epwing.or.jp/about/about.html

[42] 青空文庫, http://www.aozora.gr.jp/guide/nyuumon.html

[43] プロジェクト杉田玄白, http://www.genpaku.org/

[44] みんなの翻訳, http://trans-aid.jp/

[45] 日本語 WordNet, http://nlpwww.nict.go.jp/wn-ja/

[46] EDICT, http://www.csse.monash.edu.au/ jwb/edict.html

[池原 (2000)] 池原悟, “セマンティック・タイポロジーによる言語の等価変換と生成技術”, (平成 13 年度～18 年度：科学技術振興事業団・戦略的基礎研究 (CREST)), 2000.

[Utiyama (2003)] Masao Utiyama, et al. “Reliable Measures for Aligning Japanese-English News Articles and Sentences”, ACL-2003, pp.72-79, 2003.

[言語資源協会] 言語資源協会 (GSK), E-mail: info@gsk.or.jp, Web: http://www.gsk.or.jp/. http://www.gsk.or.jp/catalog/GSK2007-D/catalog.html

[Fujii (2010)] Atsushi Fujii, et al. , “Overview of the Patent Translation Task at the NTCIR-8 Workshop”, Proceedings of the 8th NTCIR Workshop Meeting, 2010.

[A. S. Hornby (1942)] A. S. Hornby E. V. Gatenby A. H. Wakefield, “新英英大辞典（机上版）(ISED, Idiomatic snd Syntactic English Dictionary)”, ISBN : 978-4-7589-0005-8, 1942.

[副島 (1989)] 副島隆彦, et al, “欠陥英和辞典の研究 別冊宝島”, JICC 出版, 別冊宝島 102, 雑誌 65988-55, 1989.

[副島 (1990)] 副島隆彦, Peter Van Gelder, “英語辞書大論争”, JICC 出版, 別冊宝島 113, 雑誌 65988-77, 1990.

[西山 (2005)] 西山 七絵, 村上 仁一, 徳久 雅人, 池原 悟, “単文文型パターン辞書の構築”, 言語処理学会第 11 回年次大会, pp.372-375 (2005-03).

[藤波 (2012)] 藤波 進, “これだけは知っておきたい技術者のための法律相談 法制度からみた情報流通システムの設計と運用 ”, http://www.cybersoken.com/lawlecture/index.html.

表 11　採取文数

ID	コーパス名	分類	文献	採取文数	単文	重文・複文
AA	機能試験文集	8	[1]	4,853	2,310	1,628
AB	IPAL	8	[2]	15,213	10,813	
AC	学研 アンカー和英辞典	1	[3]	46,108	14,556	14,816
AD	学研 アンカー英和辞典	1	[3]	25,274		7,011
AE	学研 英和辞典	8		4,063	1,626	846
AF	外国人のための基本語用例辞典	8	[4]	26,861	6,113	11,796
AG	三省堂 英語表現辞典	8	[5]	16,316	5,572	6,356
AH	日本経済新聞	8		6,675	952	
AI	英文ビジネスレター文例大辞典	7	[6]	12,544	1,020	5,667
AJ	外国人のための日本語例文・問題シリーズ	8	[7]	14,086	3,601	
AK	LDB	7	[8]	13,107		
AL	SENSEVAL 対訳コーパス	4	[9]	1,205	598	
AM	講談社 和英辞典	6	[10]	45,143	16,554	11,993
AN	斎藤 和英大辞典	1	[11]	94,646		19,313
AO	小倉書店 英語文型・文例辞典	7	[12]	2,133	382	835
AP	英辞郎 用例コーパス	7	[13]	16,613		3,301
AQ	研究社 新編英和活用大辞典	1	[14]	114,531	40,682	31,999
AR	ランダムハウス英語辞典	2	[15]	42,942	14,948	11,317
AS	ビジネス技術実用英語大辞典	1	[16]	13,228	2,727	5,138
AT	コンピュータ用語辞典第 3 版	1	[17]	4,067		1,705
AU	佐良木コーパス	8		1,051	20	160
AV	白井コーパス	8		1,426	43	367
AW	斎藤健太郎コーパス:比較構文	6		199	43	83
AX	澤田康子コーパス:因果関係構文	6		655	29	463
AY	アルク 英語教師用データベース	4	[18]	802	76	429
AZ	研究社 総合ビジネス英語文例事典	1	[19]	2,685	53	451
BA	新実用英語ハンドブック	1	[20]	333	126	86
BB	研究社 新和英大辞典	1	[21]	35,268	9,977	8,597
BC	機能試験文集（公開）	4	[1]	2,146		
BD	白井 2 コーパス	8		46,173		4,278
BE	三省堂 エクシード英和辞典	8	[22]	2,175	670	186
BF	科学技術日英・英日コーパス辞典	2	[23]	14,522		5,860
BG	日本語文型辞典	8	[24]	10,004	4,123	3,952
BH	旺文社マルチ辞書 辞ショック	1	[25]	64,511	27,620	
BI	田中コーパス	6	[26]	198,567		
BJ	読売新聞社説	8		560	69	
CA	アルク なるほど!英語表現データベース	4	[27]	5,997		
CB	アルク 状況別英語表現集	4	[27]	2,742		
CC	アルク 日本を紹介するキーワード	4	[27]	216		
CD	アルク カタカナ表現	4	[27]	454		
CE	アルク 四字熟語	4	[27]	300		
CF	アルク ことわざ・慣用句	4	[27]	459		
CG	アルク 擬音語・擬態語	4	[27]	327		
CH	高島康司 英文ビジネスレター	3	[28]	1,093		
CI	向井京子 英文 E メール文例集	3	[29]	1,091	264	
CJ						
CK	読売新聞 (文対応データ)	5	[30]	150,000	12,806	
CL	読売新聞 (記事対応データ)	5	[30]	5,015		
CM	ATR バイリンガル旅行会話基本構文表現一般	7	[31]	2,144		
CN	ATR バイリンガル旅行会話基本構文表現ダ文	7	[31]	608		
CO	NHK やさしいビジネス英語実用フレーズ辞典	3	[32]	7,276	773	
CP	赤尾好夫 英語基本熟語集	8				
CQ	小倉書店 自然科学系和英大辞典増補改訂新版	1	[33]	10,315		
CR	ジーニアス英和・和英辞典	1	[34]	5,394	2,330	
CS	朝日出版社 最新ビジネス英文手紙辞典 CD-ROM 版	3	[35]	2,338	176	
CT	アスク 機械を説明する英語	7	[36]	2,639	461	
ZZ	自然言語処理専門用語辞書	6				
Total				1,099,093	182,113	158,633

表 12　重文・複文の文種別

ID	コーパス名	分類	文献	重文・複文	文種別 1 (重文 1)	文種別 2 (重文 2)	文種別 3 (複文 1)	文種別 4 (複文 2)	文種別 5 (重複文)
AA	機能試験文集	8	[1]	1,628	772	55	612	54	142
AB	IPAL	8	[2]						
AC	学研 アンカー和英辞典	1	[3]	14,816	7,294	65	497	563	1,335
AD	学研 アンカー英和辞典	1	[3]	7,011	2,990	214	2,954	372	500
AE	学研 英和辞典	8		846	389	16	379	25	38
AF	外国人のための基本語用例辞典	8	[4]	11,796	6,226	1,347	2,461	344	1,420
AG	三省堂 英語表現辞典	8	[5]	6,356	3,344	310	1,964	171	570
AH	日本経済新聞	8							
AI	英文ビジネスレター文例大辞典	7	[6]	5,667	1,553	411	1,903	745	1,055
AJ	外国人のための日本語例文・問題シリーズ	8	[7]						
AK	LDB	7	[8]						
AL	SENSEVAL 対訳コーパス	4	[9]						
AM	講談社 和英辞典	6	[10]	11,993	6,343	445	4,220	292	709
AN	斎藤 和英大辞典	1	[11]	19,313	10,844	1,078	5,396	469	1,578
AO	小倉書店 英語文型・文例辞典	7	[12]	835	308	51	292	59	126
AP	英辞郎 用例コーパス	7	[13]	3,301	1,531	117	1,237	146	307
AQ	研究社 新編英和活用大辞典	1	[14]	31,999	13,153	822	14,240	1,450	2,347
AR	ランダムハウス英語辞典	2	[15]	11,317	5,505	306	4,432	328	772
AS	ビジネス技術実用英語大辞典	1	[16]	5,138	1,497	301	2,071	539	733
AT	コンピュータ用語辞典第 3 版	1	[17]	1,705	541	102	696	160	206
AU	佐良木コーパス	8		160	29	14	51	29	37
AV	白井コーパス	8		367	46	22	77	33	49
AW	斎藤健太郎コーパス:比較構文	6		83	25	14	27	7	10
AX	澤田康子コーパス:因果関係構文	6		463	323	36	46	8	51
AY	アルク 英語教師用データベース	4	[18]	429	185	56	54	30	104
AZ	研究社 総合ビジネス英語文例事典	1	[19]	451	126	35	104	77	109
BA	新実用英語ハンドブック	1	[20]	86	39	2	36	1	8
BB	研究社 新和英大辞典	1	[21]	8,597	4,590	253	2,984	172	598
BC	機能試験文集（公開）	4	[1]						
BD	白井 2 コーパス	8		4,278	2,218	32	1,929	31	83
BE	三省堂 エクシード英和辞典	8	[22]	186	118	1	62	1	9
BF	科学技術日英・英日コーパス辞典	2	[23]	5,860	1,955	562	1,692	578	1,068
BG	日本語文型辞典	8	[24]	3,952	1,950	418	976	134	476
BH	旺文社マルチ辞書 辞ショック	1	[25]						
BI	田中コーパス	6	[26]						
BJ	読売新聞社説	8							
CA	アルク なるほど!英語表現データベース	4	[27]						
CB	アルク 状況別英語表現集	4	[27]						
CC	アルク 日本を紹介するキーワード	4	[27]						
CD	アルク カタカナ表現	4	[27]						
CE	アルク 四字熟語	4	[27]						
CF	アルク ことわざ・慣用句	4	[27]						
CG	アルク 擬音語・擬態語	4	[27]						
CH	高島康司 英文ビジネスレター	3	[28]						
CI	向井京子 英文 E メール文例集	3	[29]						
CJ									
CK	読売新聞 (文対応データ)	5	[30]						
CL	読売新聞 (記事対応データ)	5	[30]						
CM	ATR バイリンガル旅行会話基本構文表現一般	7	[31]						
CN	ATR バイリンガル旅行会話基本構文表現ダ文	7	[31]						
CO	NHK やさしいビジネス英語実用フレーズ辞典	3	[32]						
CP	赤尾好夫 英語基本熟語集	8							
CQ	小倉書店 自然科学系和英大辞典増補改訂新版	1	[33]						
CR	ジーニアス英和・和英辞典	1	[34]						
CS	朝日出版社 最新ビジネス英文手紙辞典 CD-ROM 版	3	[35]						
CT	アスク 機械を説明する英語	7	[36]						
ZZ	自然言語処理専門用語辞書	6							
Total				158,633	73,894	7,085	51,392	6,818	14,440