Vol. 0 No. 0 2002

日本語英語の文対応の対訳データベースの作成

村 上 仁 一[†] 池 原 悟[††] 徳 久 雅 人[†††]

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，特に日本語 1 文が英語 1 文に対応する対訳データベースで，一般の人が入手可能なものは，存在していなかったと言って良い．本報告では，日本語 1 文が英語 1 文に対応する日英対訳データベースを，様々な電子媒体から抽出した場合の，文の量と種類について述べる．電子媒体には，CDROM，network，電子辞書など多くの種類がある．これらの努力の結果，問題は多いが，かなりの量の対訳データベースが得られることが示された．

Jin'ichi MURAKAMI,[†] Satoru IKEHARA[††]
and Masato TOKUHISA[†††]

1. はじめに

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．そして，世の中には多くの英日，日英辞書がある．しかしある程度大規模で一般の人が入手可能な日本語 1 文が英語 1 文に対応する対訳データベースは，存在していなかったと言って良い．

本報告では，日本語 1 文が英語 1 文に対応する日英対訳データベース (以後対訳データベース) を，電子辞書，CDROM，ネットワークなどの様々な媒体から抽出した場合の, 文の量と品質について述べる．

利用できる媒体としては，電子辞書，CDROM，ネットワークなどがあるが，それぞれ個別に問題がある．CDROM 版電子辞書は多くの種類があるが，対訳データベースの抽出が困難な辞書も多い．

しかし，さまざまな努力の結果，かなりの量の対訳データベースを得ることができた．本報告では，この問題と量について述べる．

† 〒 680-8552 鳥取市湖山町南 4-101
鳥取大学工学部 Tottori University
E-mail: murakami@ike.tottori-u.ac.jp
†† 〒 680-8552 鳥取市湖山町南 4-101
鳥取大学工学部 Tottori University
E-mail: ikehara@ike.tottori-u.ac.jp
††† 〒 680-8552 鳥取市湖山町南 4-101
鳥取大学工学部 Tottori University
E-mail: tokuhisa@ike.tottori-u.ac.jp

2. 使用可能な電子媒体

現在，多くの英日，日英の辞書や例文が CDROM などの電子媒体で販売されている．しかし，日本語 1 文が英語 1 文に対応した対訳データベースは，種類が限られている．しかし，電子媒体をある程度加工することで，対訳データベースを作成することが可能である．抽出可能な，電子媒体は以下の 6 つに大きく分類することができる．

(1) 電子辞書，共通 format
(2) 電子辞書，独自 format
(3) CDROM 付の書籍
(4) ネットワーク
(5) 新聞記事
(6) テキスト文，販売
(7) テキスト文，未販売
(8) その他

以後は，それぞれの特徴について述べる．

2.1 電子辞書，共通 format

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 format には，一般に公開されている format を使用している辞書と，各社独自の format 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 format には Epwing と電子ブック形式とロボワードの 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ある．日本の出版、印刷、電機、ソフトウェアに関わるさまざまな会社が協力して作った共通規約で，富

士通が中心に設定された。最初の思想は富士通のワープロ専用機 OASIS において検索する辞書という構想であったと思う．

Epwing フォーマットは 1996 年に JIS X4081「日本語電子出版検索データ構造」という JIS 規格に制定され、2001 年に拡張仕様を含めた改定が承認されている。この format の辞書は，英日，日英辞書のほかにも広辞苑，漢和辞書など，現在かなりの数が出版されている．(たぶん 50 種類を越えている．http://www.epwing.or.jp/ 参照のこと．) この format は，基本的に JIS コードで書かれている．ただし，kanji-in kanji-out がない．したがって，0x8080 の or をとれば EUC コードに変換できるため，テキストを抽出することは容易である．しかし，対訳文の抽出は，かなり困難な場合が多い．

通常，辞書に掲載されている文すべてがそのままテキストになっている．つまり，例文として掲載されている文は，テキストの中に埋め込まれている．これをなんらかのキーを基準に抽出する必要がある．対訳文が用例 file として別 file になっている辞書は，簡単に対訳文を抽出することが可能である．例として，斎藤英和大事典やビジネス/技術実用英語大辞典があげられる．しかし，このように例文を別 file として処理している辞書は少ない．また，ジーニアスなどの辞書は，単語の見出しが，例文中において〜になっている．このような辞書は，単語とキーワードとして検索するのは簡単であるが，対訳文として抽出することは困難である．

なお，電子ブックは、SONY を中心に出版・取次・印刷ソフト・ハード会社が集まって電子ブックコミッティという団体を組織し、1990 年に発売を開始したもので、当初は専用のプレイヤー（SONY の商品名はデータディスクマン）で利用する電子書籍として登場した。データはキャディーに納めた 8cmCD-ROM に収録されており、プレイヤーにはキーボードと液晶（モノクロ２階調）画面が備わっていた。「広辞苑」、「現代用語の基礎知識」、「模範六法」などをはじめとする多数のタイトルが出ている。1991 年 10 月には International Electronic Book Publisher's Committee (IEBPC) が組織され、海外でもタイトルが出版されるようになった．現在電子ブックになっている製品の多くは Epwing 形式でも出版されている．

ロボワードは株式会社テクノクラフトが開発発展させてきた OS に依存しない辞書検索システムである．このシステムは，コンピュータとのインターフェースが良いため，使い勝手が良いシステムになっている．そのため，かなりの辞書がこのシステムで検索できる．しかし，辞書は，圧縮がかかっていることや外字登録がされているため、解析は困難を伴う．そのため，現状ではこの format の辞書から対訳データベースを抽出することは行っていない．

2.2 電子辞書，独自 format

電子辞書では独自の format をとり，専用のブラウザでなければ見えないものがある．これらの多くは windows のみで動作する．これらを解析することは非常に手間隙のかかる作業である．しかし，ランダムハウス英語辞典は，歴史のある辞書であるためか．format を解析して Epwing 形式に変換する tool が web に掲載されている．これを利用することで対訳文が抽出できる．また，ビジネス技術実用英語大辞典は用例が多いため解析する価値があると判断し，format を調べて，解析し，対訳例文を抽出した．

ただし，すべての辞書に対して format を調べて解析することは困難である．特に辞書に圧縮や外字がある場合，解析は困難である．

2.3 CDROM 付の書籍

最近 CDROM 付の書籍が販売されている．この中から，日英の対訳のある本を探して，簡単なスクリプトをつくることで，対訳文が抽出可能である．今後，対訳データベースを作成する良い素材になると思う．ただし，1 冊のなかに大量の対訳文は載っていない．

例として，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 2000 がある．

2.4 ネットワーク

ネットワーク上に公開されている対訳データベースがある．基本的には，中学校の英語の文法を教えるための定型文で，アルク社が多い．ただし，一定の時期にしか公開されていなく，現在は見れない．例として，英語教師用データベースがある．(http://www.alc.co.jp/ 参照)

2.5 新 聞 記 事

新聞記事は，言語リソースにおいて非常に重要なものである．大量のテキストデータが必要な場合，新聞記事がもっとも入手しやすい．

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 eveing news，読売新聞と The Daily Yomiuri，毎日新聞と Mainichi Daily News がある．これらの中から，読売新聞から出版されている The Daily Yomiuri が個別に CDROM で購入できる．しかし，記事対応にすらなっていないため，対訳文の作成には非常に困難である．

この対応を自動的に見つける研究があり，それなり

の対応がとれた対訳データベースが作成されている．

ただし，このデータベースは非常に高価である．

2.6 テキスト文，販売

わずかな例ではあるが，英日対訳でテキスト文が販売されている．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があげられる．

2.7 テキスト文，未販売

英日対訳でテキスト文があるが，基本的には，個人もしくは会社で収集した対訳データベースである．翻訳の研究のために作成されたものであるため， 対訳文としては最適なものが多い．しかし，一般の人には入手不可能であるのが残念である．

例として IPAL の英文がある．IPAL の日本語文を，NTT が翻訳実験のために翻訳した対訳データベースである．この場合，英語の著作権は NTT にあると思われる．

2.8 そ の 他

現在 network 上において，青空文庫などのように著作権が切れた本を掲載するプロジェクトが発足している．

http://www.aozora.gr.jp/guide/nyuumon.html

これらの本に対して翻訳をするプロジェクトがある．(プロジェクト杉田玄白，http://www.genpaku.org/) このプロジェクトを利用することで本の対訳文が得られるが，残念ながら 1 文ごとの対応にはなっていないため，対訳データベースの抽出は困難である．

また，各外国語大学には，それぞれ独自の対訳データベースがあるようだが，全容は不明である．

3. 抽出したデータベース

以下に各データベースごとの概略を述べる．

3.1 機能試験文集

分類は 4.

現鳥取大学の池原悟教授が NTT に在籍のころ作成したもの．機械翻訳システム評価用に使用している文例集である．3,718 文は以下のところから入手可能である．

http://www.kecl.ntt.co.jp/icl/mtg/resources/index-j.html

ただし，オリジナルは 5240 文であり，全文は公開されていない．公開されている 3,718 文は池原コーパスとして登録している．

詳細は，以下の論文を参照のこと．

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)*

3.2 IPAL

分類は 8.

IPAL のなかの『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』の日本文を翻訳者に委託して翻訳した対訳文である．著作権は NTT にあると思われる．一般の入手は不可能と思われる．

紹介文を以下に示す．

英和約 *51,000* 語を収録した *IPAL* は，特別認可法人 情報処理振興事業協会 *(Information-technology Promotion Agency, Japan)* では *1981* 年 *10* 月の技術センター発足当初から，情報処理分野における日本語処理の重要性に着目し，日本語辞書 *(IPALexicon of the Japanese Language for computers)* の研究・作成に取り組んでまいりました．

IPAL には，語彙体系上ならびに使用頻度上重要であると考えられる基本的な動詞 *(861* 語*)*， 形容詞 *(136* 語*)*， 名詞 *(1,081* 語*)* について，意味および統語的な特徴に基づいて区分し，それを一つの単位として，形態，意味，統語，慣用表現などに関する情報が詳細に記載されています．

http://www.ipa.go.jp/STC/NIHONGO/IPAL/ipal.html

3.3 アンカー和英辞典 アンカー英和辞典

分類は 1.

別形態として，NTT が辞書から人手で入力した対訳データベースがある．また，電子ブック版もある．

紹介文を以下に示す．

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* 語を収録した『ニューアンカー和英辞典』の *CD-ROM* です．英語の熟語や例文を含まれている英単語を組み合わせることにより探すことができます．日常語を重視し，英作文，英会話にすぐに役立つ便利な文型表示も用意してあります．なお，本製品は検索ソフトおよび辞書データをハードディスクにインストールしてご利用いただくことができます．その場合 *4MB*，および *47MB* の空き容量が必要です．

定価 *6,300* 円 *[*税別*]*

■お問い合わせ先 *045-475-9*

3.4 学研英和辞典

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を NTT が辞書から人手で入力した対訳文．この本は現在販売されていない．

3.5 基本語用例辞典

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，NTT が翻訳者を使用して翻訳し

た対訳データベース．原文には英語がない．著作権は NTT にあると思われる．一般の入手は不可能と思われる．

紹介文を以下に示す．

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
初版：*1994* 年／第 *3* 版：*2000* 年 出版社：大蔵省印刷局
A5 判 *1,337P 4,854* 円 *4-17-151302-2*

3.6 英語表現辞典

分類は 8.

NTT において，本から人手によって打ち込んだ対訳データベース．(と思っている．)

紹介文を以下に示す．

三省堂編修所 編
2,100(2,000) 円 B６変 *736* 頁 *4-385-11012-3*
基本的な日本語のキーワードのもとに，その語に関連するいろいろな表現を幅広く集め，それぞれに適切な英語訳を与え，日本語表現に応じた豊かな英語表現ができるようにした和英。本編では代表的動詞を収録。巻末に個々の表現からも引ける詳しい索引付き。
1997 年 *5* 月 *1* 日 発行

3.7 日本経済新聞

分類は 8.

NTT において，翻訳した対訳データベース．新聞記事のため，日本語の文が非常に長い．そのため英語も長いものになっている．とても機械翻訳ではできそうにない文である．

3.8 英文ビジネスレター文例大辞典

分類は 6.

英日対訳でテキスト文が販売されている，希少な例である．

言語学研究に携わる皆様へ 『英文ビジネスレター文例大辞典 *CD-ROM* 版』を，下記のフォーマットで言語学研究のために特別にご提供します．共通言語資源として，ぜひご活用下さい．
媒体 *CD-ROM* データフォーマット形式
1. 日本語と英語のデータを別々のファイルに格納，それぞれのファイルは行単位で対応になります．
2. データは *SGML* 形式
3. 文字形式は *S-JIS*

発行元 日本経済新聞社 発行日 *1998* 年 *2* 月 納品 申込書到着後 *10* 日以内 価格 *70,000* 円 *(*税別・送料込み*)*
ご購入にあたっては，言語学研究用使用許諾契約書の締結が必要です．本ページから *PDF* ファイルで提供する「言語学研究用使用許諾契約書」を *3* 部印刷し，必要事項記入捺印のうえ，下記宛先へ郵送お願いします．お申し込み書はファクシミリでも承りますが，発送は上記許諾契約書の受領後になります．

お申し込み先 〒 *103-0025* 東京都中央区日本橋茅場町 *1-6-10*
株式会社日経出版販売映像ソフト部行
TEL:03-5651-3725 FAX:0120-21-0082
E-Mail:eizo@nikkeish.co.jp

http://www.nikkeish.co.jp/gengo/eibun.htm

3.9 外国人のための日本語例文・問題シリーズ

分類は 8.

NTT が翻訳者を使用して翻訳した対訳データベース．原文には英語がない．著作権は NTT にあると思われる．一般の入手は不可能と思われる．

3.10 LDB

分類は 6.

ATR が作成した自動音声認識のためのホテル対話などを収録したデータベース．基本的には，音声とテキスト両者がある．

このデータベースは，現在も収録が続けられていて，内部には 20 万対話のデータになっている (らしい．) ATR では，用例翻訳，統計翻訳のための対訳データベースになっている．

3.11 SENSEVAL 対訳コーパス

分類は 6.

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問題である．このために，対訳データベースを作成した．多くは単語の訳であるが，一部，文章に対する英訳がつけられている．

3.12 講談社和英辞典

分類は 8.

電総研で辞書から人手によって入力されたデータである．文字誤りが多い．岐阜大学，池田先生より入手した．電子技術総合研究所から入手することが可能である．

---

o 電総研によって電子化された和英辞典．対訳例文約 *38,000* 文を含む．
o 研究目的に限る．使用のための誓約書を電総研と交わす必要がある．
連絡先: 元吉文男 〒 *305* 茨城県つくば市梅園 *1-1-4*
電子技術総合研究所 知能情報部 自然言語研究室
Tel:(0298)58-5914 Fax:(0298)58-5930
Email:moto@etl. go. jp
http://cactus.aist-nara.ac.jp/lab/resource/resource-print.html#KODANSHA

3.13 斉藤和英大辞典

分類は 1.

1938 年に斎藤秀三郎が出版した辞書．A5 版、見出

し語 5 万、用例 12 万、総頁数 4640 頁という、当時としては前例のない大和英辞典であった。斎藤はその序文において、「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．

辞書の中に用例が別 file としてあるので，対訳文が簡単に抽出できる．しかし，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文がある．翻訳文も意訳が多いため，賛否わかれる辞書である．

例文
山が崩れても動かぬ (女は尻が重い)
She will not budge an inch.

なお，斎藤秀三郎は昭和 4 年に 64 歳でなくなっているため，この辞書の著作権は切れていると考えられる．そこで，この辞書にはいっている例文を人手によって修正したものを公開する予定である．

英語辞典史上の金字塔「斎藤和英大辞典」*(*昭和 *3* 年，日英社刊*)* を，現代仮名遣いに改めるなどしてデジタル化．古き良き日本語表現を多く含む約 *5* 万の見出しと約 *15* 万の文例・用例を様々な方法で検索することができ，対訳表現辞典としても使えます．

ISBN4-8169-8078-4
定価 *18,000* 円 *[*税別*]*
■お問い合わせ先 *03-3763-5241*

3.14 小倉書店 英語文型・文例辞典

分類は 6.

英日対訳でテキスト文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 文章である．

下記の内容について，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もので， 作文作業中に必要な項目を立ち上げ，目的の文書に適う用例を活用すれば，この分野の慣用的な表現と文書の骨組が決まります． つまり，「論文の書き方」といったハウツウ式の参考書と首っ引きで自己流の作文をするのとは， 手間と正確さが違います．姉妹編である『自然科学系和英大辞典』を枝葉とすればその幹となるものです．

http://www.ogurashoten.co.jp/kyozai3.html

3.15 英辞朗 用例コーパス

分類は 6.

nifty でネットワーク上で作成した，単語数では最大の辞書で，100 万単語を越える．単語数と比較すると例文が非常にすくないのが惜しまれる．テキスト形式で配布されているため，対訳文を作成するのは簡単である．

紹介文を以下に示す．

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データの名称です．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*)* などが含まれています．

英辞郎は本来，翻訳・通訳の仕事に必要なものとして個人的趣味で作っていたものですが，パソコン通信で知り合った人たちが「私にも使わせて」とおっしゃるので，*@nifty* の英会話フォーラムのデータライブラリを通じて無料で配布しております．

3.16 研究社 新編英和活用大辞典

分類は 1.

例文は比較的分かりやすい法則で 1 行に収まっているのでほぼすべてが抽出できたと思われる．ただ、公表されている用例数とかなりの差がある．

紹介文を以下に示す．

「英語を書くための大辞典」として各界で英語を使う人々に好評の『新編英和活用大辞典』をそっくり収録．名詞を中心に語と語の慣用的な結びつき *(*コロケーション*)* を徹底的に採録し，その用例総数は *38* 万．見出し語検索のほか，すべての用例中の任意の英単語からの自由語検索や，見出し語の訳語からの簡易和英検索機能などを搭載．

ISBN4-7674-3573-0
定価 *13,000* 円 *[*税別*]*
■お問い合わせ先 *03-3288-7777*

3.17 ランダムハウス英語辞典

分類は 2.

独自の format 形式．しかし，Epwing の format に変換する情報がネットワークにあるため，抽出はさほど困難ではない．英文は，非常によいが、日本語は英文を翻訳して作成しているため，日本語に奇異な文があるらしい．

本文は main.txt に暗号化して格納されている．暗号といっても 0xff で xor を取れば簡単に復号化できる．本文は SJIS で格納されている．また、本文中で使用されている外字を ASCII 文字で置き換える変換テーブルは太田純氏の CSRD というユーティリティー付属の CSRD 外字代替表記ファイルを流用した．

紹介文を以下に示す．

派生語などを含めると収録語数 *34* 万 *5,000* 語．この膨大な項目を簡単に検索できる何種類もの検索方法を用意しました．しかも発音記号のつく約 *12* 万語はネイティブスピーカーによる発音を聴けるようになっているなど，マルチメディア英語辞典ならではの使いやすさです．さらに書籍版にはなかった和英機能を新たに追加．英語研究者，翻訳者などから，英語を学ぶ人，ビジネスで使う人まで，英語にたずさわるすべての人たちに必携のデジタル辞典です．

http://ebook.shogakukan.co.jp/scatalog/randam/top/top.htm

3.18 ビジネス技術実用英語大辞典

分類は 1.

Epwing 形式，用例が別 file になっているため，非常に抽出が楽.

個人的には，この例文がとても美しい (無駄がない) と感じている.

紹介文を以下に示す.

海野文男+海野和子 *[編]* 価格 *11，000* 円 *(*税別*) ISBN4-8169-8127-6 T4937695181270 2000* 年 *12* 月発売

収録数大幅 *UP!* 実用性の高い用例が *15* 万件■最新データを追加し，内容さらに充実*!* 見出し語数が前版の英和 *17，000* 語，和英 *18，000* 語から各々 *19，000* 語，*20，000* 語へ，用例数が *126，000* 件から *150，000* 件へと大幅アップ.

英語圏で実際に使われている生きた表現

実際に英語圏で使用された文書や印刷物をもとに，*15* 万件を超える“ 生きた用例・文例を抽出し，整理・編集した辞典です．用例は全てネイティブによって書かれた，実用性の高い表現を精選．既存の辞書では調べのつかない重要表現も多数収録．豊富な用例が英文作成/英語翻訳に威力を発揮します.

http://www.nichigai.co.jp/newhp/whats/unno3.html

3.19 コンピュータ用語辞典第 3 版

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ので、今回は英文の文末に.(ピリオド) がきちんと付いている和英の例文のみを抽出した.

紹介文を以下に示す.

---

*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〔発行/発売〕*2000* 年 *12* 月発売

ISBN4-8169-8126-8 T4937695181263 標準価格 *11,000* 円 *(*税別*)*

『*CD-*コンピュータ用語辞典 第 *2* 版』*(1997* 年 *5* 月発売*)* の改訂新版*!* 情報処理，データ通信・ネットワーク，情報理論，*OA*，パソコンなどコンピュータ関連分野を対象に，基本的な用語から話題の最新用語まで，英和 *33,000* 語，和英 *35,000* 語，用例・文例 *12,600* 件を収録した座右の用語集．重要語にはすべて解説，用例・文例付き.

http://www.nichigai.co.jp/newhp/whats/computer3.html

3.20 佐良木コーパス

分類は 8.

佐良木さんが個人的に収集した対訳データベース

3.21 白井コーパス

分類は 8.

井さんが個人的に収集した対訳データベース

3.22 鳥取大学池原研究室 斎藤健太郎コーパス:比較構文

分類は 8.

鳥取大学工学部知能情報学科池原研究室の斎藤健太郎さんが多くの本から，比較構文のみを収集した対訳データベース

3.23 鳥取大学池原研究室 澤田康子コーパス:因果関係構文

分類は 8.

鳥取大学工学部知能情報学科池原研究室の澤田康子さんが因果関係構文のみを多くの本から収集した対訳データベース

3.24 英語教師用データベース

分類は 4.

アルクが公開している英語教師用の対訳データベース．ただし，現在公開が中止されている.

紹介文を以下に示す.

「英語教師用データベース」は，学校や英語サークル，塾などで英語を教えている人のためのページです．このページで提供している英語学習素材は，高等学校の英語の授業で必修となっている英語レベルを想定して作成しています．次の *3* つのメニューがあります．現在は,【必修構文 使える例文集】のみ公開しておりますが，他 *2* メニューも近日オープン予定です.

1. 【必修構文 使える例文集】*:*高等学校で必修の英語構文を使った例文を収録．例文集のテキストはダウンロードも可 *(*←公開中*!)*

2. 【機能別フレーズ集】*:*「悲しい」「うれしい」などの英語の機能をもとに分類したフレーズ集．*2002* 年度から採用される，文部省の英語教育指導要領に完全準拠 *(*←近日オープン予定*)*

3. 【教室で使える英語フレーズ】*:*教師が実際の授業中に使うであろうフレーズを英訳 *(*←近日オープン予定*)*

http://home.alc.co.jp/db/owa/engt_structure?stg=4

3.25 研究社 総合ビジネス英語文例事典

分類は 1.

この辞書では例文が複数行にわたっているため、抽出プログラムはやや複雑になる．しかし，プログラムから見て分かりやすい例文はほぼすべて取り出すことができた。

紹介文を以下に示す.

研究社ビジネス英語スーパーパック

研究社

日常業務から情報収集，海外向け発信まで，ビジネスのあらゆる場面に役立つビジネス英語ツールを集大成．ビジネス関係の諸分野の専門用語を幅広く収録した『研究社ビジネス英和辞典』，ビジネス現場のさまざまなシーンで使われる英語のサンプルレターと豊富な活用文例約 *7000* を収録した『総合ビジネス英語文例事典』，*26* 万語を収録した実務家に定評のある『リーダーズ英和辞典』*(*初版*)* を *1* 枚の *CD-ROM* にスーパーパック.

ISBN4-7674-3590-0

定価 16,000 円 [税別]
お問い合わせ先 03-3288-7777

3.26 新実用英語ハンドブック

分類は 1.

この辞書では例文が複数行にわたっているため、抽出プログラムはやや複雑になる．しかし，プログラムから見て分かりやすい例文はほぼすべて取り出すことができた。

紹介文は以下に示す.

実務に必要な英語を集大成した便利な事典「実用英語ハンドブック」の最新版を，図表も含め全文収録した CD-ROM. 貿易通信，輸出入書式から契約，為替実務，会社の定款，国際データ通信，広告・カタログ，スピーチまで扱い，ビジネスレター作成を始め，英語によるビジネスに威力を発揮する．検索ソフト (Windows，PowerMac 用) を添付した．

ISBN4-469-74233-3
定価 7,700 円 [税別]
お問い合わせ先 03-3295-6231

3.27 研究社 新和英大辞典

分類は 1.

この辞典も例文は比較的分かりやすい法則で 1 行に収まっているのでほぼすべてが抜き出せたと思われる．なぜか、公表されている例文数は 5 万なのに、実際に抜き出してみると約 20 万になった.

合成語・句は 16 万、例文 5 万と書かれていて、合計すると 21 万になるので、おそらく合成語・句・例文を抜き出したと思われます。

紹介文を以下に示す.

最大規模の収録語彙を誇り，国内のみならず海外でも評価高い『新和英大辞典』の CD-ROM 版．印刷辞書のすべて (見出し語 8 万，合成語・句 16 万，例文 5 万) に， 時事用語・コンピューター用語など 6200 語を増補．見出し語はローマ字・かな・漢字表記のいずれでも検索可能.

ISBN 4-7674-7200-8
定価 12,000 円 [税別]
お問い合わせ先 03-3288-7777

3.28 エクシード英和辞典

分類は 8.

白井さんが収集したデータベース

紹介文を以下に示す.

三省堂編修所 編
1,995(1,900) 円 B7 変 992 頁 4-385-10650-9
類書中最多の 12 万項目収録。コンピュータ・バイオテクノロジー・ビジネス・企業名・地名・人名など現代生活のあらゆる分野から選定。俗語・口語・略語・複合語・成句も充実。簡潔ながら十分な訳語。探し易い配列。

1998 年 11 月 10 日

3.29 科学技術日英・英日コーパス辞典

分類は 2.

独自の format 形式．今回抽出した辞書のなかでもっとも情報が少ない．windows のコードをハッキングして適当なコードから対訳を抽出した． 片山氏の手腕に依存した．用例が多い.

紹介文を以下に示す.

本体 18,000 円+税 ISBN4-621-04991-7 丸善
●著者が長年にわたって科学技術・産業実務関係の書籍・雑誌・パンフレットなどから収集した約 30 万点にもおよぶ膨大な英例文の中から，精選された 6000 のパラグラフ (文節)，センテンス数にして約 15000 文例を収録.
●約 15000 の英語例文すべてに日本語対訳付き.
http://pub.maruzen.co.jp/cd_others/ko-pas/

3.30 佐良木 2 コーパス

分類は 6.

新聞記事の社説.

読売新聞の社説において，The Daily Yomiuri と比較して 1 文 1 文人手によって抽出した対訳データベース

3.31 旺文社 マルチ辞書 辞ショック

分類は 1.

紹介文を以下に示す.

信頼の旺文社辞書「ニューサンライズ英和辞典」「ニューサンライズ和英辞典」を収録．収録語数約 7 万 6 千，例文・イディオム数約 12 万．見出し語の前方・後方一致検索はもちろん，例文やイディオムの検索，画面語検索など多彩な検索が可能．混乱しやすい語の使い分けを示す図やイラストも豊富で，英作文や英文和訳に大活躍の一枚です.

定価 5,800 円 [税別]
お問い合わせ先 03-3267-6865 株式会社アスク

3.32 田中コーパス

分類は 6.

兵庫大学の田中康仁さんが，学生に作成させた対訳文日本語から英訳した文．学生が作成したため，品質が低いかもしれないとのこと．対訳文の量は，かなりある．

参考論文を以下に示す.

言語処理学会 第 8 回 年次大会 (NLP2002) プログラム
B4-2 日英・パラレルコーパスの作成
○田中康仁 (兵庫大学)

3.33 ア ル ク

分類は 4．以下の 6 つに分類される．

アルク Oh，that's how you say it! なるほど! 英語表現データベース

アルク Situation 100 plus すぐに使える! 状況別英語表現集

アルク 日本を紹介するキーワード

アルク カタカナ表現

アルク 四字熟語

アルク ことわざ・慣用句

アルク 擬音語・擬態語

ただし，現在公開が中止されている．

紹介文は以下に示す．

「*SPACE ALC*」では，オンライン辞書の決定版ともいえる英和・和英辞書「英辞郎 *on the Web*」を始め,「ビジネス英語辞書」「語源辞典」などの充実した辞書を無料で提供しております．この中にあって「英語表現事典」は，日本の日常生活におけるさ まざまな事象「最新の流行表現」「日本独特の言い回し」などを英語で表現している，数少ないコンテンツです．従来の辞書等では調べることが困難だった表現をカバー し，たいへん好評をいただいております．また「英語表現事典」には，すべての語に例文とコンパクトにまとめられた解説が掲載してありますので，知りたい言葉の意味を見るだけではなく，その言葉の使い方まで理解できるようになっています．

【英語表現事典 *INDEX*】

◆ *Oh, that's how you say it!* なるほど*!* 英語表現データベース*:*ふだんなにげなく口にするひとことを集めた英語表現のデータベース．投稿者から寄せられた日常のひとことを，随時データベースに蓄積しています．

◆ *Situation 100 plus* すぐに使える*!* 状況別英語表現集*:*日常表現を，生活の中の場面毎に整理した状況別英語表現集．

▽英語ならこう言う*!*シリーズ▽

◆日本を紹介するキーワード*:*「フリーター」「出会い系サイト」など，日本独特の文化や社会現象について外国人に説明するとき，必ず押えておきたいキーワードを集めた会話表現集です．

◆カタカナ表現*:*アンケートは*?*ナイターは*?*クレームは*?*私たちが日ごろ口にするカタカナ表現は，果たしてそのまま英語として通じるのでしょうか.「英語ならこう言う」が分かるカタカナ英語表現集です．

◆四字熟語*:*「意気投合」「四苦八苦」「自業自得」など，たった四文字で微妙な ニュアンスを表現する四字熟語．日本語ならではの表現を英語に変換した四字熟語英語表現集です．

◆ことわざ・慣用句*:*「焼け石に水」「馬子にも衣装」など，日本語の表現を豊かにしていることわざや慣用句．英語ならではの面白い言い回しも豊富に紹介する，ことわざ・慣用句英語表現集です．

◆擬音語・擬態語*:*「ピューピュー」,「ゴロゴロ」,「キーキー」など物事の状態や様子を表現する擬音語や擬態語．こういった表現を用いて日ごろ伝え合っている微妙なニュアンスを，英語で伝えるとしたら*?*ことばの幅がぐっと広がる，擬音語・擬態語英語表現集です．

3.34 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 2000

分類は 3．広告文を以下に示す．

高水準の"確実に通じる"英文が書ける．*CD-ROM* には全文を収録．英文ビジネスレター実用フォーマットと例文集

CD-ROM 付 英文ビジネスレター実用フォーマットと例文集
高島 康司 *4-939076-25-3 527* ページ *A5* 判 *1,900*
2000 年 *2* 月 *25* 日

さまざまなケースに使える豊富な実用フォーマットと，組み合わせが自由自在のたくさんの文例で，さまざまな用件に対応したビジネスレターが，すぐに正確に書くことができます．本の内容を素早く検索でき，フォーマットと例文をコピーして使える *CD-ROM* が付いて，時間短縮に役立てられます．

3.35 向井京子 英文 E メール文例集 池田書店 2002

分類は 3.

紹介文を以下に示す．

世界の人とコミュニケーション*!* プライベートでもビジネスでも役立つ *E* メール英語表現集 *1000*

著者 向井京子体裁 *A5* 判*/256* 頁定価 本体価格 *1600* 円*+*税
ISBN4-262-16896-4

グループ旅行や家族旅行，一人旅や留学など海外に出かける日本人は年々増え続け，国際交流も豊かになってきました．帰国後，そんな旅行先でお世話になった現地のガイドさんや友人，ホームステイでお世話になったホストファミリーなどとのコミュニケーションに役立つのがこの「英文 *E* メール文例集」です.「ありがとう．とってもたのしかったです」「いま日本はとっても寒いんです」などメールは手紙とは違って簡便なツールですので気軽るにコミュニケーションすることができます．またインターネット上で知り合った友人とも「おしゃべり」できるよう事項紹介などの項目も盛り込みました．巻末には意外と難しい件名 *(subject)* のインデックスや *ASAP* や *BBL* などといた略語なども収録，そしてなんといっても *CDROM* つきなのでコピー*&*ペストでスムーズに文章が作成できます．ぜひ一度書店さんでご覧ください *(*編集担当*:*田口*)*.

3.36 読売新聞記事

読売新聞には英字新聞として The Daily Yomiuri がある．これらは，別々に販売されている．そして，記事対応にすらなっていない．しかし，最近 CRL において日本語 1 文に対して英語 1 文に対応つけたデータベースが，読売新聞の記事を購入することを前提に，入手可能になっている．

この新聞記事に対して，日本語 1 文に対して英語 1 文に対応づける研究を以下にしめす．

第 *151* 回 自然言語処理研究会、第 *68* 回 情報学基礎研究会（合同開催）

3. 日英新聞記事の対応付けと精度評価内山将夫， 井佐原均 *(*通信総合研究所*)*

読売新聞と *The Daily Yomiuri* との記事対応および文対応を得たので，その手法および精度を報告する．

ただし，読売新聞の DB は非常に高価である．1 年分の日本語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．

紹介文を以下に示す．

言語処理学会員各位

読売新聞の記事データが研究目的で利用できるようになりましたので，お知らせいたします．利用できるようになったデータは以下の通りです．

1987 年以降の邦文新聞記事 *1989* 年以降の英文新聞記事邦文新聞記事に付与されているキーワードの元になる「ヨミダス用語辞書」

詳細は，*http://www.ndk.co.jp/yomiuri/* をご覧下さい．

3.37 抽出不能なデータベース

例文抽出が困難な CDROM(Epwing) をしめす．

(1) ジーニアス英和・和英辞典

分類は 1.

例文の中で単語や連語は〜で略されているため，仮に例文を取り出せたとしても肝心の単語を対応づけることが困難である．つまり，文中で省略されている単語を補完するのは困難である．

紹介文を以下に示す．

CD-ROM 版 ジーニアス英和・和英辞典
大修館書店
好評の「ジーニアス英和辞典¡改訂版¿」*(9* 万 *2* 千語句収録*)* と，ユニークな¡ハイブリッド方式¿による「ジーニアス和英辞典」*(8* 万語句収録*)* を *1* 枚に収録．英語を入力すると英和辞典が，日本語 *(*漢字も可*)* を入力すると和英辞典が引ける．英和では，指定の語を含むイディオム・用例を直接検索可能．音声 *1* 万 *4* 千語収録．*ISBN4-469-79057-5*
定価 *9,000* 円 *[*税別*]*
お問い合わせ先 *03-3295-6231*

---

(2) リーダーズ英和辞典 研究社

分類は 1.

例文の中で単語や連語は〜で略されているため，仮に例文を取り出せたとしても省略されている単語を補完するのは困難である．

Epwing 版 リーダーズ+プラス *V2*
研究社
『リーダーズ英和辞典』*(*第 *2* 版*)* と『リーダーズ・プラス』の合わせて *46* 万語の全文字データを収録した最新・最大の英和辞典．一般語をはじめ，固有名・専門語・新語・俗語・イディオム・略語など情報化時代に要求されるあらゆる項目を貪欲に取り込んだ実務家向け英和辞典の決定版．和英辞典としても活用できる訳語検索，綴りのわからない単語も引けるカタカナ検索 *(*約 *6* 万語*)* など機能満載．*ISBN4-7674-3563-3*
定価 *20,000* 円 *[*税別*]*
お問い合わせ先 *03-3288-7777*

---

(3) ニューセンチュリー英和・新クラウン和英辞典

分類は 1.

外字が多い．

紹介文を以下に示す．

新グローバル&ニューセンチュリー英和・和英辞典 *ISBN*：*4-385-61400-8*
本体 *6,500* 円+税
1999 年 *4* 月 *30* 日発行
JAN コード：*T4938641614002*
木原研三・小西友七ほか 編

書籍版『新グローバル英和辞典』、『ニューセンチュリー和英辞典』は、ご好評をいただき、今日までたくさんの読者の皆様に支えられてまいりました。英和辞典は収録語数 *93,000*、収録成句 *13,000* 余、和英辞典は収録語数 *41,000* を数えます。全用例は *15* 万例におよび、学習辞典として高い評価を得ています。

---

(4) ＣＤ－科学技術４５万語対訳辞典　英和・和英

分類は 1.

例文がない．

紹介文を以下に示す．

発行・発売：日外アソシエーツ
価格 *38,000* 円（税別）
ISBN4-8169-8128-4　　*2001* 年 *7* 月 *20* 日発売

基礎科学（科学技術一般、数学、物理学ほか）をはじめ、電気・電子・情報、医学、生物学、化学、機械、地球科学など幅広い分野の専門用語、約 *45* 万語（日本語訳）を *CD-ROM1* 枚に収録。
技術の進歩が著しい情報・通信分野の用語をはじめ、最新の *JIS* 用語などを大幅追加。
見出し（英語・日本語）のもとに、対訳や略語のほか、その用語が主に使用される分野名を表示。また、最新の *JIS* 用語については、規格番号と正式名称も表示しています。
語意、綴りが簡単に確認でき、科学技術関連の技術文書やマニュアルなどの翻訳、論文発表、調査・研究の際の基本的なツールとして、またビジネスや学習用途にもご利用いただけます。
弊社『*EB* 科学技術用語大辞典』（*1991* 年発売）に最新の用語、*JIS* 用語等を追加収録したものです。

---

(5) 研究社　新英和・和英中辞典

分類は 1.

１文１ファイルらしき状態で抜きだせている．外字があるものも存在する．根性を入れれば抜き出せる．

紹介文を以下に示す．

研究社 新英和・和英中辞典辞書&検索ソフトシリーズ研究社 新英和・和英中辞典
価格 *6,800* 円（税別）
定評ある『新英和中辞典』『新和英中辞典』の最新版の全文字情報を収録した改訂新版。総収録語数は英和 *9* 万、和英 *7* 万、用例数は英和 *8* 万 *3* 千、和英 *10* 万。書籍を切り替えることなく英和・和英を自在に検索。用例・成句の直接検索などの便利な検索機能に加え、英和の見出し語約 *1* 万 *5* 千語にネイティブによる音声データを付加。学習や指導に役立つ各種の単語リストを付録として収録。なお、本製品は検索ソフトおよび辞書データをハードディスクにインストール

してご利用いただくことができます。その場合、*6MB* および *91MB* の空き容量が必要です。

---

4. 得られた対訳データベースの量

4.1 抽出した対訳データの文数

抽出できた対訳データの文数を表 1 中の「抽出文数」に示す．この数字は，文章と推定して抽出した文の数であるため，名詞や名詞句も抽出されている．

4.2 抽出した重文，複文の文数

抽出できた対訳データから，人手によって文を整形し，重文および複文を抽出した．ただし，4.1 で機械的に抽出したデータベースすべてに対して，重文，複文の抽出を行ってはいない．抽出した重文，複文の文数を表表 1 中の「重文複文」に示す．

4.3 文 種 別

抽出した，重文，複文を以下の文種別にしたがって分類した．これらを表 1 中の文種別 1 から 5 に示す．

(1) 文種別１
文接続１箇所を持つ文
例文
私は椅子に座り、窓の外を眺めた。
彼の声は大きいので、部屋中に響いた。
自動ドアの前に立つと、ドアが開く。
テレビを見ながら勉強してはいけません。
自宅に電話をかけたが、誰も出なかった。

(2) 文種別２
文接続２箇所を持つ文
例文
明日は東京へ行って、友達に会い、久しぶりに話をする。
エアコンはないし、部屋は狭いし、この物件はあまり良くない。
電話をしながらメモを取ったが、そのメモをなくしてしまった。
昨日バナナを買ったけれど、まだ青くて食べられないだろう。

(3) 文種別３
埋め込み文１つを含む文
例文
さっきまでここにいた猫がいなくなった。
私たちが数学を習った先生に街で偶然会った。
お客様にお渡しする品物は、丁寧に扱ってください。
先週渋谷で見た映画をもう一度見たい。
声の大きい人を募集しています。
昨日より楽しい話をするつもりだ。
空がきれいな日に海へ行きたい。

(4) 文種別４
埋め込み文２つを含む文
例文
約２０名を乗せた飛行機が人の住んでいない島に不時着した。
私が手紙を書いた人は以前友人の紹介で知り合った人だ。
昨日私が買い物をしたデパートにナイフを持った男が立てこもっているらしい。
父と撮った写真をアメリカにいる姉に送った。
娘より若い女性が私の落とした財布を拾ってくれた。

(5) 文種別５
文接続１箇所＋埋め込み文１つの文
例文
彼女の吐き出す息は白く、頬は赤かった。
母が作ってくれたマフラーを見ると、学生時代を思い出す。
兄は父が設立した会社を手伝いながら、こつこつとお金を貯めている。
彼は親友が通っている大学へは行かず、他の大学を選んだ。
今までより安い光熱費にしたいなら、節約するしかない。
店内が静かな喫茶店に入り、コーヒーを飲んだ。
肉が嫌いな人がいたら、代わりに魚を頼みます。

（注）埋込について修飾要素を持たない用言（連体形）は埋め込み文とはしない。

5. 謝 辞

この報告書を書くには，多くの人の協力がありました．具体的には，ATR 音声翻訳通信研究所 第三研究室 室長の白井諭さんと，長崎純心大の佐良木 昌さんには，個人的に収集して頂いた対訳データベースを頂きました．電子辞書からの対訳文の抽出には，鳥取大学，知能情報工学科 2 年の片山 慶一郎氏の助力を得ました．また，

データベースの文種別の分類には NTT-AT の，木村淳子氏，小見佳恵氏、阿部さつき氏、村本奈央氏、小船園望氏が中心になって行いました．これらの方に感謝いたします．

表 1 抽出した文数

データベース名	抽出文数	重文，複文	文種別 1	文種別 2	文種別 3	文種別 4	文種別 5
機能試験文集	5,240						
IPAL	30,707						
アンカー和英辞典	46,108	14,816	7,294	65	497	563	1,335
アンカー英和辞典	25,278	7,011	2,990	214	2,954	372	500
学研英和辞典	4,063	846	389	16	379	25	38
基本語用例辞典	28,896	11,796	6,226	1,347	2,461	344	1,420
英語表現辞典	16,393	6,356	3,344	310	1,964	171	570
日本経済新聞	10,000						
英文ビジネスレター文例大辞典	15,087	5,667	1,553	411	1,903	745	1,055
外国人のための日本語例文・問題シリーズ	25,028						
LDB	32,769						
SENSEVAL 対訳コーパス	6,920						
講談社和英辞典	57,927	11,993	6,343	445	4,220	292	709
斉藤和英大辞典	149,113	19,313	10,844	1,078	5,396	469	1,578
小倉書店 英語文型・文例辞典	3,779	835	308	51	292	59	126
英辞朗 用例コーパス	17,301	3,301	1,531	117	1,237	146	307
研究社 新編英和活用大辞典	281,215	31,999	13,153	822	14,240	1,450	2,347
ランダムハウス英語辞典	137,696	11,317	5,505	306	4,432	328	772
ビジネス技術実用英語大辞典	63,847	5,138	1,497	301	2,071	539	733
コンピュータ用語辞典第 3 版	6,451	1,705	541	102	696	160	206
佐良木コーパス	1,118	227	1,010	523	1,611	674	1140
白井コーパス	1,668	367	46	22	77	33	49
斎藤健太郎コーパス:比較構文	242	83	25	14	27	7	10
澤田康子コーパス:因果関係構文	639	463	323	36	46	8	51
英語教師用データベース	802	429	185	56	54	30	104
研究社 総合ビジネス英語文例事典	2,703	451	126	35	104	77	109
新実用英語ハンドブック	333	86	39	2	36	1	8
研究社 新和英大辞典	191,795	8,597	4,590	253	2,984	172	598
池原コーパス	3,718	1,628	772	55	612	54	142
白井 2 コーパス	4,278	4,278	2,218	32	1,929	31	83
エクシード英和辞典	2,175	186 1	118	1	62	1	9
科学技術日英・英日コーパス辞典	14,627	5,860	1,955	562	1,692	578	1,068
佐良木 2 コーパス	851						
旺文社 マルチ辞書 辞ショック	106,945						
田中コーパス	211,997						
アルク なるほど! 英語表現データベース	5,997						
アルク 状況別英語表現集	2,742						
アルク 日本を紹介するキーワード	216						
アルク カタカナ表現	454						
アルク 四字熟語	300						
アルク ことわざ・慣用句	459						
アルク 擬音語・擬態語	327						
高島康司 英文ビジネスレター	1,093						
向井京子 英文 E メール文例集	1,609						
読売新聞 (文対応データ)	150,000						
読売新聞 (記事対応データ)	30,000						