Journal of Culture and Information Science, **1**(1), 17－36. (March 2006)

研究ノート

Developing “Resources for Corpus Linguistics”

Kenji Kitao

This article and the resource provided here introduce the reader to corpus linguistics resources available on the web, through “Resources for Corpus Linguistics,” which the author has developed with English and Japanese annotations. Following an explanation of corpus linguistics and its usefulness, the author explains how the resource is organized and what is included in each section. “Resources for Corpus Linguistics” is included in this article. The original resource which is frequently updated can be found at
http://www.cis.doshisha.ac.jp/kkitao/Japanese/library/resource/corpus/corpus.htm.

1. Introduction

1.1 Why corpus linguistics?

Corpus Linguistics is a relatively new field in linguistics and a new approach to languages. “Corpus” was defined in *The Oxford English Dictionary* in 1956 as “The body of written or spoken material upon which a linguistics analysis is based” Later the term “corpus” came to be used to refer to machine readable texts, and they are called “computer corpora” or “electronic corpora.”

Corpus linguistics using computers started with the Brown Corpus compiled by Kucera & Francis, which was compiled between 1961 and 1964 at Brown University in the United States. It has fifteen different text categories with 500 2000-word texts written in 1961, with a total of one million words (Leech, 1991). The results of this study were published by Kucera & Francis (1967), which formed the foundation the practice of making large electronic corpora and analyzing them by computer.

The term “corpus linguistics,” has been used frequently since Aarts & Meijs (1984), and it became well established by around 1990, which was the end of its first stage in this field. During this period, Lancaster-Oslo/Bergen Corpus (LOB) was compiled. It is a counterpart of Brown Corpus with British English in 1961 (one million words) (1970-1978). It was grammatically tagged (all words were given a word-class label) (1978-1983). While LOB Corpus was compiled, Brown Corpus was tagged (1970-1978). These taggings were parts of speech, and they contributed to studies of frequency of parts of speech (Francis & Kucera, 1982; Johansson & Hofland, 1989).

Starting in 1991, the British National Corpus (BNC) was compiled by Oxford University Press and six other organizations (1991-1994), with 90 million written words and 10 million spoken words with tags. This was opened to the European Union in 1995 and to the world in 2000. The Bank of English (BoE) was compiled by Harper-Collins and Birmingham University (1991-1995) with 200 million words for the purpose of compiling dictionaries and studying English. The BoE is still being added to, and there are more than 550 million words. These huge corpora have been used to compile dictionaries. A counterpart of BNC with American English is being compiled as American National Corpus (ANC). All these corpora are tagged for parts of speech for vocabulary and grammar analyses.

Since the middle of 1980s, specialized corpora have been compiled such as ones for Australian English, New Zealand English, Old English, Middle English, Modern English, etc., for studying individually or comparing different Englishes. Parallel corpora, corpora composed of the same texts in two different languages, have been compiled specifically for comparing two languages.

In order to process many and large corpora, concordancers, software to search in corpora and

display the results so that they can be analyzed in terms of vocabulary, grammar, and other aspects of language, were developed. Most of these tools are available for free or at a nominal fee. They are easily and cheaply available for individual users for the study of language. This makes studying corpora possible for people, such as language students and teachers, who may not otherwise have the technical knowledge or funds to do so.

Computers and the Internet have developed greatly over the past two decades. Individuals can afford to use computers and the Internet, and personal computers are faster and can store more linguistic data than even large computers 20 or 30 years ago. There are many free texts which are already electronic corpora on the Internet. There are also many CD-ROMs which have electronic texts. It is easy to make personal corpora for the purpose of analyzing it. Many corpora, including the BNC, the Brown Corpus, and the LOB are available on the Internet, and anyone can use them without charge.

Thus almost all researchers, teachers of English language, and students who study English or other languages can use and take advantage of corpora already made and concordancers that are available online. It is easy to make a corpus, tag it, and analyze it.

1.2 What can corpus linguistics do?

Before corpus linguistics, linguists studied grammar based on authority or native speaker intuition. However, with corpora, linguists and teachers or students of language can use actual linguistic data to show how language works. They can analyze vocabulary in terms of frequency, collocations, and grammar. They can test descriptions of grammar in a grammar book. They can describe the language and show qualitative and quantitative analyses. They can use corpora as a large dictionary or a grammar book. They can compile corpora of specific types of language and find the frequently used vocabulary and expressions as well as some grammar rules in that area of language. They can compare corpora of the speech and writing of learners of English with corpora of standard English and find out what types of errors and patterns of errors those learners make. There are many ways to look at the language using corpora.

1.3 Why "Resources for Corpus Linguistics"?

Corpus linguistics has developed greatly over the past ten or fifteen years. There are many resources available on the Internet, ranging from very technical to very basic, provided by researchers and teachers of English. Some are not comprehensive and some are not specific as to how they were compiled. There is no comprehensive resource for beginners of corpus linguistics, English language teachers or students who might use concordances and corpora for language teaching or studying the English language.

The author has developed this site, "Resources for Corpus Linguistics," mainly for researchers of English, English teachers, and students. He has gathered as much information as possible, except for highly technical resources that ordinary researchers probably would not use.

The author has classified sites into the following categories: "General resources (including lists of links)", "Online dictionaries," "Vocabulary frequency checkers," "Vocabulary frequency lists," "Search engines," "Concordancers," "Online corpora," "Tagging," "Parallel corpora," "Learner corpora," "Training for corpus linguistics," "Analyzing language using corpora," and "Corpora available on the Internet (mainly non-copyrighted)." The author has put brief annotations for each link so that readers can understand what type of resource to expect.

2. What is "Resources for Corpus Linguistics?"

This resource includes useful links related to corpus linguistics, with annotations. It has thirteen sections. (The original resource is available at http://www.cis.doshisha.ac.jp/kkitao/Japanese/library/resource/corpus/corpus.htm and is updated frequently by the author.)

The following is a description of each section and what is included.

2.1 General resources

This section includes web sites which have general resources of corpus linguistics, including lists of links.

2.2 Online dictionaries

This section includes mainly English dictionaries, English-Japanese dictionaries, and Japanese-English dictionaries. They can be used to find the meanings of

specific words and find sample sentences using the word, so that users can understand how those words are used in context.

There are links to dictionaries, thesauruses, and specialized dictionaries.

2.3 Vocabulary frequency level checkers

There are several sites that have vocabulary frequency level checkers, where users enter an English passage to find out the frequency level for each word in the passage, using a vocabulary frequency list. This section includes the JACET8000 word checker, an index often used in the college level English teaching.

2.4 Vocabulary frequency lists

This section has links to the sites of different vocabulary frequency lists. These lists are useful for determining vocabulary levels for teaching materials or for different levels of students.

2.5 Search engines

This section includes search engines such as Google and the sites related to using Google for corpus linguistics.

2.6 Concordancers

This section includes sites which have information on concordancers and sometimes on corpora. There are also web-based or downloadable concordancers.

2.7 Online corpora

This section includes the sites where users can search corpora using a web based concordancer. This section includes well known corpora such as the BNC, the LOB, and the Brown Corpus.

2.8 Tagging

This section includes the sites where users can get tagging software and web sites where they can cut and paste English passages to be tagged.

2.9 Parallel corpora

This section includes Prof. Asao and Prof. Uchiyama’s sites, where users can experience how parallel corpora work using newspapers, passages, etc.

2.10 Learner corpora

This section includes the sites which have links and information about learner corpora or a particular learner corpus. Well known corpora are included here.

2.11 Training for corpus linguistics

This section includes the sites where users can learn about corpus linguistics, how to use some concordancers, how to make their own corpus, using Excell and editors for corpus analysis, finding regular expressions, and statistics. Using the links in this section, users can understand the basics of corpus linguistics and learn to use concordancers effectively.

2.12 Analyzing language using corpora

This section has the sites where users can change formats and/or analyze corpora and the explanations for using those sites.

2.13 Electronic texts (non-copyrighted corpora)

This section includes many non-copyrighted electronic texts. A very few of them may still be under copyright. Users can use them when they make their own corpora.

3. How to Use “Resources for Corpus Linguistics” Effectively

This resource is designed for corpus linguistics researchers, English language teachers, and students. It can be used by people who do not know anything about corpus linguistics or even by experts.

3.1 For beginners and students

The author strongly encourages users to read whatever material appears useful in the section “Training for corpus linguistics” first. Some sites have comprehensive information about corpus linguistics.

Some sites provide information for using a certain corpus or a concordancer. Some have information about making a corpus and doing linguistic analyses. For Japanese-speaking people, there are some comprehensive handouts from workshops.

For the actual linguistic analyses, “Analyzing language using corpora” is a very useful section, particularly for Japanese-speaking people. Those sites explain how to organize corpora for analysis, how to analyze them, and what corpus linguistics can actually do.

After finishing the useful sites in those two sections, the author suggests that users explore web sites where certain corpora can be searched in the “online corpora”

section The users will become aware of BNC, Brown Corpus, LOB, and many other corpora available online.

After getting acquainted with online corpora, the author suggests users explore "Concordancers." The users will develop further knowledge of concordancers and how to use them effectively. They can download concordancers and use them as a linguistic analysis tool.

The author suggests skimming the "General resources" section, since it has a vast amount of information about corpus linguistics. However, it is a good idea to stick to what is necessary and useful for the user, otherwise too much information that might not be useful will be obtained.

If the user is interested in online dictionaries and vocabulary study, the "online dictionary" section is very useful. There are a vast number of dictionaries available online. Some special dictionaries are useful for certain subject areas.

If the user is interested in vocabulary frequency, "vocabulary frequency level checkers" and "vocabulary frequency lists" are useful. Using the former, words in the passage are divided into different levels of frequency. The latter explains different frequency lists.

If the user is interested in making his/her own corpus, "Electronic texts (non-copyrighted corpora)" provides a vast number of texts for study. Since most of them are not copyrighted, the user can show the actual corpus in his/her research papers.

3.2 For users who know corpus linguistics

Those who are already familiar with corpus linguistics can choose any section which might be useful for them. "General resources" might be very informative for anyone, and it is worth skimming it.

The sections "Online corpora," "Training for corpus linguistics," and "Analyzing language using corpora" are likely to be particularly useful, and it is worth skimming the sections and trying links that sound interesting. The latter two are useful for corpus studies using a personal corpus. Resources in "Electronic texts (non-copyrighted corpora)" are useful for making personal corpus.

3.3 For English teachers

This resource is useful for English teachers. They can take advantage of resources in the sections "Online dictionaries," "Vocabulary frequency level checkers," "Vocabulary frequency lists," "Concordancers," and "Online corpora." These are useful for making materials or preparing for classes. Resources in "Parallel corpora" might be useful to introduce a new way of teaching English. Teachers can learn a great deal about the types of errors students make and use the information to improve their teaching by studying resources in "Learner corpora."

4. Resources for Corpus Linguistics

(The original site of this resource is at http://www.cis.doshisha.ac.jp/kkitao/Japanese/library/resource/corpus/corpus.htm.)

4.1 General resources

●Bookmarks for Corpus-based Linguists (by Dr. David Lee)
http://devoted.to/corpora
**A well-organized searchable and browsable list of about 1000 links related to corpus linguistics. Categories include English corpora; non-English corpora; software, tools, and frequency lists; and references, papers, and journals. There are explanations of the web page and of each category. Extensively annotated.

●ICAME (International Computer Archive of Modern and Medieval English)
http://nora.hd.uib.no/icame.html
**The web page of the International Computer Archive of Modern and Medieval English, whose purpose is to "collect and distribute information on English language material available for computer processing and on linguistic research completed or in progress on the material, to compile an archive of English text corpora in machine-readable form, and to make material available to research institutions." It includes the ICAME journal, a list of links, corpus manuals, and online use of the ICAME corpus (for registered users of the ICAME CD-ROM). Unannotated.

●Links to corpus linguistics & related sites (by Przemek Kaszubski)
http://www.staff.amu.edu.pl/~przemka/corplink.html
**Extensive list of links, some annotated, including such categories as downloadable software for corpus work, corpora and language teaching, and lexical and lexicographic resources.

●Prof. Tono's site

http://leo.meikai.ac.jp/%7Etono/
**Annotated list of links, with emphasis on learner corpora.
●Phrases in English
http://pie.usna.edu/
**Software for various kinds of searches of the British National Corpus, with explanations.
●ELISA
http://www.uni-tuebingen.de/elisa/html/elisa_index.html
**A corpus of spoken English from interviews with native English speakers, allowing users to either browse or search the corpus.
●Corpora4Learning.net
http://corpora4learning.net/
**A bibliography, English corpora, and tools and websites, with extensive annotations.
●Gateway to Corpus Linguistics
http://www.corpus-linguistics.info/
**Extensively annotated links to corpora-related resources, corpora, and so on.
●ICT4LT
http://www.ict4lt.org/en/index.htm
**Continually updated training modules related to information and communications technology.
●The Complete Lexical Tutor
http://www.lextutor.ca/
**A resource for data-driven learning, with tutorials on lexical development, annotated links to resources for teachers.
●Corpus Linguistics
http://www.rc.kyushu-u.ac.jp/~higuchi/text7/corpus.html
**Briefly annotated links to corpora, software, and other resources.
●Centre for English Corpus Linguistics
http://cecl.fltr.ucl.ac.be/
**Information about corpus linguistics, including information about conferences and other events, a bibliography, publications, etc. Unannotated.
●Corpus Linguistics
http://www.staff.amu.edu.pl/~przemka/
**List of links with some brief annotations.
●Corpus Linguistics
http://www.sfb441.uni-tuebingen.de/c1/corp-ling-engl.html
**Briefly annotated list of links to corpora in a wide variety of languages, software, and other resources.
●Corpus linguistics, translation, and language learning
http://www.federicozanettin.net/sslmit/cl.htm
**Briefly annotated list of links to corpus linguistics organizations, software, conferences, etc.
●Tim Johns Data-Driven Learning page
http://www.ecml.at/projects/voll/our_resources/graz_2002/ddrivenlrning/whatisddl/resources/tim_ddl_learning_page.htm
**Links with good explanations to various resources, including Tim Johns’ concordance-based teaching and learning materials, a bibliography, various software, and other links pages.
●Michael Barlow's Corpus Linguistics site
http://www.athel.com/corpus.html
**Annotated links to corpora in more than 20 languages, learner corpora, and so on.
●UCREL Home Page (Lancaster University)
http://www.comp.lancs.ac.uk/computing/research/ucrel/
**This is the web page of a research center for corpus linguistics. It includes extensively annotated links to papers on the subject, to corpora, to search tools, to tagging tools, etc.
●Someya's Homepage
http://www.cl.aoyama.ac.jp/~someya/
**A list of links related to online concordancers, English for business purposes, interpretation studies, etc. (Pull-down the menu labeled “Choose and Click.”)
●Corpus resources available in the linguistics lab
http://bulba.sdsu.edu/corpus-resources.html
**Annotated links to corpora in English, Japanese, and Chinese.
●Croatian Language Technologies
http://www.hnk.ffzg.hr/jthj/corpora.htm
**Links to corpora in Croatian as well as about 20 other languages, dictionaries, corpus-related tools, information about conferences, etc.
●Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources
http://nlp.stanford.edu/links/statnlp.html
**Annotated list of links to corpora, tools, syllabi, etc.
●19 Corpus Linguistics
http://www.essex.ac.uk/linguistics/clmt/other_sites/index_19.html
**Uncategorized links to corpus-related resources, a few of them annotated.
●Corpus Linguistics and Written Language Resources Bibliography
http://liceu.uab.es/~joaquim/language_resources/lang_res/biblio_corpus.html
**An extensive list of books, articles, and other

resources related to corpus linguistics.
●Corpus Analysis- Online Resources
http://www3.telus.net/public/mcleanky/bridgingthegap/corpresources.html
**A well-annotated list of links.
●Corpus Linguistics
http://plaza.snu.ac.kr/~hskwon/corpus.html
**An unannotated list of links and textbooks related to corpus linguistics.
●Corpus Linguistics Links Guide.
http://links-guide.ru/sprachen/linguistik/corpus-linguistics.html
**A few annotated links related to corpus linguistics.
●Corpus Resources
http://pioneer.chula.ac.th/~awirote/ling/corpuslst.htm#Collocation
**A short, unannotated list of links related to collocations.
●TESL : Linguistics
http://iteslj.org/links/TESL/Linguistics/
**An annotated list of links related to lisnguistics.
●English Corpora I
http://corp.hum.ou.dk/itwebsite/corpora/corp/page3.html
**List of links to corpora, categorized according to type of English.
●Corpora
http://www-user.tu-chemnitz.de/~voigt/link_corpora.htm
**An annotated list of links to a variety of corpora, divided by type.

(Sites in Japanese)

●Prof. Nishino's links
http://muse.doshisha.ac.jp/corpus/index.html
**Prof. Nishinoh's extensive and useful list of links related to corpus linguistics is divided into categories, including dictionaries, search engines, electronic texts, online libraries, and searchable online corpuses. Partly in Japanese. Unnannotated.
●Prof. Goto's corpus linguistics and natural language processes links
http://www.sal.tohoku.ac.jp/~gothit/textprocessing.html
**Unannotated list of links, mostly in Japanese.
●On-Line Corpus（2004/08/03）
http://www.kct.ne.jp/~takaie/CORPUS%20site.htm
**Unannotated. Mostly in Japanese.
●List of pearlscripts and CGI
http://oscar.gsid.nagoya-u.ac.jp/program/perl/
**List of links, unannotated and mostly in Japanese
●Prof. Inoue's site
http://lexis.ias.tokushima-u.ac.jp/
**List of links in Japanese
●Mr. Takaie's site
http://www.kct.ne.jp/%7Etakaie/
**List of links, mostly in Japanese, some with annotations.
●Corpus linguistics links
http://lexis.ias.tokushima-u.ac.jp/link/corpus.html
**Unannotated links, some in Japanese, to corpora, concordance programs, etc.

4.2 Online dictionaries

●Terminology Collection
http://www.uwasa.fi/termino/collect/index.html
**WORD-ONLINE General Language Dictionaries ・ TERM-ONLINE Special Language Glossaries
●OneLook Dictionary Search
http://www.onelook.com/
**A site where numerous dictionaries can be searched.
●Merrian Webster OnLine
http://www.m-w.com/
**Searchable dictionary and thesaurus, plus other resources, such as words of the day, downloads, etc.
●The American Heritage Dictionary of the English Language
http://www.bartleby.com/61/
**Searchable dictionary that also includes information about regional words, English as a living language, usage notes, and Indo-European roots.
●Reference Materials for Students and Researchers
http://www1.doshisha.ac.jp/~kkitao/online/www/referenc.htm
**List of links to reference materials, writing resources, style sheets, search engines, etc., mostly briefly annotated.
●WordNet 2.0 Vocabulary Helper
http://poets.notredame.ac.jp/cgi-bin/wn
**WordNet 2.0 Vocabulary Helper is an interface to the WordNet-2.0 lexical database. Word meanings are organized in a semantic network. This interface produces a lot of output for each word.
●Roget's Thesaurus of English Words and Phrases
http://poets.notredame.ac.jp/Roget/
**A reference book that was written in 1911 showing English words and phrases organized by concept.
●Spelling checker for English words
http://poets.notredame.ac.jp/cgi-bin/spell
**A spelling checker with interface to International

Ispell using uses an American English dictionary.
●Longman Dictionary of Contemporary English Online
http://www.ldoceonline.com/
**A free online version of the CD-ROM that comes with the dictionary.
●Oxford Advanced Learner's Dictionary (OALD)
http://www.oup.com/elt/catalogue/teachersites/oald7/?cc=global
**A free online dictionary.
●Oxford English Dictionary (OED)
http://www.oed.com/
**A premium online dictionary.
●Middle English Dictionary
http://ets.umdl.umich.edu/m/med/
**For authorized users only. (premium site)
●Britannica Online
http://www.britannica.com/
***Encyclopedia Britannica*, with a premium area with access to the full encyclopedia, a free area with “concise” encyclopedia entries, plus such features as “Biography of the Day” and “This Day in History.” (premium site)
●Lexical Resources
http://www-a2k.is.tokushima-u.ac.jp/member/kita/NLP/lex.html
**Large, partially annotated list of links to a variety of types of resources related to words, including dictionaries, thesauruses, word frequency lists, etc.
●ROGET'S Thesaurus Search Form
http://humanities.uchicago.edu/orgs/ARTFL/forms_unrest/ROGET.html
**A search form for Roget’s 1911 thesaurus.

(Sites in Japanese)

●ARIADNE GLOSSARIES アリアドネ 辞書・事典・用語集
http://ariadne.ne.jp/glossary.html
**一般・総合 （百科事典・国語辞典・人名事典・辞書類のリンク集など）法律・経済・ビジネス用語エネルギー・通信用語・言語学用語 （英語辞典・類語辞典・略語辞典・俗語辞典・諺・常套句辞典）多言語間翻訳ツール・双方向辞典｜英和・和英｜英仏・仏英・仏仏｜文学用語芸術用語 （音楽・美術・舞踊など）
●Internet Watch
http://internet.watch.impress.co.jp/static/link/2004/05/21/jisho.htm
**国語、英和、和英辞典と翻訳サービス
●オンライン辞書・辞典リンク集
http://www.hir-net.com/link/dic/
**国語辞典 英和 和英 英英 他言語 医学 コンピューター インターネット
●辞書のリンク
http://www.kotoba.ne.jp/
**英和･和英･国語・英英・百科事典･自動翻訳
●オンライン辞書 English Navi
http://koho.ktplan.jp/link008.html
**オンライン辞書のリンク集
●辞書･事典 (Dictionaries)
http://www1.doshisha.ac.jp/~kkitao/japanese/online/dictionary.htm
**リンク集 / マルチ言語辞典 / 英和辞典 / 和英辞典 / 学習者辞典 / 英英辞典 シソーラス / 百科事典 / 子供向け辞書 / 日本語の辞典 / その他の辞書・事典
●ＯＮ-Line Dictionary 一覧
http://www.kct.ne.jp/~takaie/dictionary.htm
**Partially annotated list of online dictionaries, mostly in Japanese.
●Infoseek 翻訳
http://www.infoseek.co.jp/Honyaku?pg=honyaku_top.html
**Translator of texts among English and Japanese, Japanese and Chinese, and Japanese and Korean.
●英次郎
http://www.eijiro.jp/
**英和･和英
●英辞郎
http://www.alc.co.jp/index.html
**英和･和英
●翻訳から生まれた新世代の英和辞典
http://corpus.idd.tamabi.ac.jp/yourei/index.htm
**英和
●WebGrep for NESS 6800
http://cow.gsid.nagoya-u.ac.jp/program/webgrep/webgrepNESS.html
**杉浦研究室が収集した英語学習用英日対訳例文集に例文６８００
●WebGrep (名古屋大学 杉浦研究室)
http://cow.lang.nagoya-u.ac.jp/program/webgrep/
**用例も豊富
●Excite 翻訳
http://www.excite.co.jp/world/english/
**英日・日英の両方が可能
●Infoseek 辞典
http://jiten.www.infoseek.co.jp/Eiwa?pg=jiten_etop.html&col=EW

**Searchable English-Japanese, Japanese-English, and Japanese-Japanese dictionary.
●Goo 辞書
http://dictionary.goo.ne.jp/index.html?&kind=ej
**英和・和英・国語
●Infoseek マルチ辞書
http://jiten.www.infoseek.co.jp/Eiwa?pg=jiten_etop.html&col=EW
**英和・和英・国語

4.3 Vocabulary frequency level checkers

●Frequency Level Checker
http://language.tiu.ac.jp/flc/
**A tool that tells users how frequent the words are that appear in a text (3 levels of frequency, first 1000, second 1000, and third 800, plus "other" and numbers). Produces a color-coded text, list of the number of total, types, and families for each level, as well as percentages of types and families at each level.
●Checker
http://language.tiu.ac.jp/flc/tool.html
**This is the checker itself from above. Users can specify which color each level should appear in, or whether it should be invisible.
●Web Frequency Indexer (Georgetown)
http://www.georgetown.edu/faculty/ballc/webtools/web_freqs.html
**A tool that produces a list of words based on how frequently they appear in the text, or alternatively all the words in the text in alphabetical order.
●JACET 8000 LEVEL MARKER
http://www01.tcp-ip.or.jp/~shin/J8LevelMarker/j8lm.cgi
**A tool that tells users how frequent the words are that appear in a text (8 levels of frequency, 1000 words each, plus "other"). Produces a text with color-coded words.
●Eva Text Analysis
http://poets.notredame.ac.jp/cgi-bin/evatext
**Input a text in the Text Entry Form and select output options. Eva Text Analysis can add markup for the WordNet 2.0 Vocabulary Helper, check spelling, and show word frequencies.
●The AWL Highlighter
http://www.nottingham.ac.uk/~alzsh3/acvocab/awlhighlighter.htm
**A program where users can copy and paste a text and have core academic vocabulary identified, based on the Academic Word List. Users can choose the level of vocabulary to be identified.
●Academic Vocabulary
http://www.nottingham.ac.uk/~alzsh3/acvocab/index.htm
**An explanation of the academic vocabulary list used in the above Highlighter.

(Sites in Japanese)

●Gakugei University TEFL
http://www.u-gakugei.ac.jp/~tefldpt/misc/goi/level.html
**教材に含まれる語彙レベルを測定するソフト、WEB 上で処理可能。ダウンロードして使用も可能
●「JACET 8000 語のチェック」の使用方法の解説と例
http://www.cis.doshisha.ac.jp/kkitao/Japanese/library/resource/corpus/j8lm.cgi.htm
**JACET 8000 LEVEL MARKER の使用方法

4.4 Vocabulary frequency lists

●ESL:Vocabulary: Lists
http://iteslj.org/links/ESL/Vocabulary/Lists/
**Annotated list of links to different types of word lists, including the 100 most commonly used English words; 5000 collegiate words, with definitions; and lists of common adjectives and adverbs.
●The First 100 Most Commonly Used English Words
http://www.duboislc.org/EducationWatch/First100Words.html
**A word list with numbers but with no definitions.
●300 Most Commonly Used Words
http://www.myenglishlessons.net/most_common.htm
**Word list in frequency order with no definitions and no numbers.
●3000 Most Commonly Used Words in the English Language (USA)
http://www.paulnoll.com/China/Teach/English-3000-common-words.html
**Word list based on magazines and newspapers with numbers but no definitions. Includes an explanation of how the list was compiled.
●5000 Collegiate Words with Brief Definitions
http://freevocabulary.com/
**Word list in alphabetical order. Intended for SAT preparation, but not necessarily the most frequent or common.
●The 6,318 Most Commonly-Used Words in English?
http://www.ie.reitaku-u.ac.jp/~provo/index5.htm
**Explanation of how the list was compiled and how it is organized, with links to a printable list or zip files.
●The Academic Word List

http://language.massey.ac.nz/staff/awl/awlinfo.shtml
**List based on families, arranged in alphabetical order with an indication of frequency (10 frequency levels of 60 families each) and by frequency (excluding the 2000 most frequent words in English).
●The Academic Word List
http://www.nottingham.ac.uk/~alzsh3/acvocab/wordlists.htm
**An academic word list arranged by frequency with links to exercises and an explanation of how to use concordencers; a general service list of about 2300 words.
●BNC database and word frequency lists
http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html
**Two frequency lists based on the British National Corpus.
●Word Frequencies in Written and Spoken English: based on the British National Corpus.
http://www.comp.lancs.ac.uk/ucrel/bncfreq/
**A companion web site to the book.
●General Service List 2000 words
http://jbauman.com/aboutgsl.html
**The General Service List from 1953 (2000 words selected as of most “general service” to learners of English), with explanations.
●OGDEN's BASIC ENGLISH
http://ogden.basic-english.org/basiceng.html
**A list of 850 words (with some supplementary lists) that cover 90% of the concepts covered by the 25,000 words in the Oxford Pocket English Dictionary.
●University Word List 808
http://jbauman.com/aboutUWL.html
**A list of 808 words that are common in academic texts, divided into 11 levels.
●Words and Phrases
http://www.eslgold.com/vocabulary/words_phrases.html
**Various lists of words and phrases, divided by level and type.

(Sites in Japanese)

●Hokkaido University English Vocabulary List
http://icarus.ilcs.hokudai.ac.jp/jugyo/huvl/
**北海道大学言語文化部英語系で作られた、大学生向けの英語語彙表
●語彙リスト（2004/01/05）
http://www.kct.ne.jp/~takaie/vocabulary.htm
**Unannotated list of links to vocabulary lists, mostly in Japanese.
●名大基本語 5000
http://www.lang.nagoya-u.ac.jp/~tonoike/linda5000.html
**JACET が作った 4000 語のリスト、また Longman の辞書で記述用に使う単語としてリストされているものなどを参考に、Linda Woo が作成
●杉浦リスト　高校 5776 語
http://grad.nufs.ac.jp/sugiura/sugiuralist.html
**この語彙リストは、文部省検定教科書『英語 I』（平成 12 年度用）48 種類の使用語彙のうち、頻度２以上の 5,776 語の頻度（全体で何度現れたか）とレンジ（何種類の教科書に現れたか）と表示されている。
●アルク 12000 語
http://www.alc.co.jp/goi/index.html
**アルクの作成した 12000 語の語彙頻度表

4.5 Search engines

●AltaVista
http://www.altavista.com
**A search engine that also has machine translation software.
●Google
http://www.google.com
**One of the most widely used search engines.
●Google Directory
http://directory.google.com/Top/Reference/Thesauri/
**List of links to thesauri, with annotations.
●open directory project
http://www.dmoz.org/
**Annotated and organized list of links to a wide variety of subjects.

(Sites in Japanese)

●Google による用例検索
http://infosys.gsid.nagoya-u.ac.jp/~ohna/ggl/
**Google を有効に使用した用例検索方法
●Google 検索エンジンを使用した表現チェック
http://www1.doshisha.ac.jp/~kkitao/japanese/library/call/google.doc
**Google で表現を検索する、簡易な説明
●Google を使用して、KWIC 表示
http://163.136.182.112/xyz01/
**佐藤先生の GoogleFormatter で、Google 検索結果 100 例までを KWIC 表示できる。使用方法は下の解説参照
●GoogleFormatter
http://www.soc.hyogo-u.ac.jp/tani/GForm.html

**An explanation of how to use GoogleFormatter as a KWIC search engine.

4.6 Concordancers

●Concordancing
http://www.nsknet.or.jp/~peterr-s/concordancing/
**A web page for concordancing.
●ICAME CORPUS MANUALS
http://khnt.hit.uib.no/icame/manuals/index.htm
**Links to corpora and concordancers
●KWIC Finder
http://www.kwicfinder.com/KWiCFinder.html
**A page from which users can download a KWIC search program with explanation, advice, links to papers and presentations, etc.
●Tools & websites
http://corpora4learning.net/resources/materials.html
**Links to manuals to a variety of corpora, using Adobe Acrobat.
●Cambridge International Corpus
http://www.cambridge.org/elt/corpus/cic.htm
**Information about the corpus, which has over 700 million words, including both spoken and written British English, American English, and a learner corpus. Can only be used by writers and authors for Cambridge University Press.
●A Ten-step Introduction to Concordancing through the Collins Cobuild Corpus Concordance Sampler
http://web.quick.cz/jaedth/Introduction%20to%20CCS.htm
**Explanation of how to use the concordancing program with the Collins Cobuild Corpus Concordance Sampler.
●Frame Net
http://framenet.icsi.berkeley.edu/index.php
**"The Berkeley FrameNet project is a lexicon-building effort for contemporary English in which we (1) select words with particular meanings; (2) describe the frames or conceptual structures which underlie these meanings; (3) examine sentences containing these words. as found in a very large corpus of contemporary English; and (4) record the ways in which the components of the sentences containing these words express information about the frames they evoke."
●Frame Net を使用した論文
http://sato.fm.senshu-u.ac.jp/_web/papers/_frmsql.pdf
**佐藤弘明先生の解説
●MICASE (Michigan Corpus of Academic Spoken English)
http://www.hti.umich.edu/m/micase/
**A browsable or searchable corpus of nearly two million words made up of transcripts of academic speech events.
●Spaceless.com
http://www.spaceless.com/concord/
**A concordancer where users can use html files as text to analyze.
●WebKWIC
http://prairie.lang.nagoya-u.ac.jp/program/webkwic-e.html
**A concordancer where users can cut and paste passages to analyze.
●Virtual Language Centre Web Concordancer.
http://vlc.polyu.edu.hk/concordance/
**A concordancer that allows users to search English, Chinese, French, Japanese, or parallel texts.
●ConcApp Concordancing Programs
http://www.edict.com.hk/PUB/concapp/
**A downloadable concordancing program that includes "full editing support and testing activities, and also word frequency text analysis" which can process most European languages and Chinese, Russian, Thai, and Japanese.
●How to Use ConcApp
http://www.edict.com.hk/PUB/concapp/Help/tutorial1.HTM
**A tutorial for using ConcApp.
●Simple Concordance Program
http://www.textworld.com/scp
**A concordance program that allows you to search text in various languages (and add alphabets) for words, phrases, and patterns.
●miniAPPolis.com Downloads
http://miniappolis.com/downloads.html
**Free downloads of various programs useful for corpus linguistics.
●TextWorld.com
http://www.textworld.com/
**Two free downloads, a concordancing program and a collocation finding program.
●Child Language Data Exchange System
http://childes.psy.cmu.edu/
**A system that provides tools for studying conversation, including transcripts, tools for coding, etc

(Sites in Japanese)

●【コーパス言語学 - 主なコーパス】
http://lexis.ias.tokushima-u.ac.jp/corpus_ling/corpus02.html

**Links to corpora with explanations in Japanese.
●English Phrase Collection
http://epc.hp.infoseek.co.jp/
**映画のせりふ 17 表現を検索、17 万語
●杉浦研究室で開発したプログラム
http://oscar.gsid.nagoya-u.ac.jp/program/
**杉浦研究室で開発された Java, JavaScript, Perl, CGI などのプログラム
●KWIC Concordance for Windows
http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/kwic.html
**単語リスト、インデックス、コンコーダンス、コロケーションリストを得るソフト
●KWIC Concordance for Windows の解説
http://www.babel.co.jp/mtsg/corpus/kwic/kwic.htm
**ソフトの使用の解説
●CAT (Corpus Analysis Toolkit)
http://www.eonet.ne.jp/~lago/cat/index.htm
**perl のプログラムで、単語頻度 、n-gram 頻度、grep、KWIC コンコーダンス、コロケーション（頻度順）、コロケーション（統計値）を求められます。
●TXTANA
http://www.biwa.ne.jp/~aka-san/
**TXTANA をはじめとするコーパス・言語分析のシェアウェア・フリーウェアを公開
●Treebank Search
http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/treebank_search.html
**Penn-Helsinki Parsed コーパス用検索ソフトウェア　単語を指定して、コーパスから検索をする。その際構文解析された情報を同時に取り出すことができる。

4.7 Online corpora

●VIEW: VARIATION IN ENGLISH WORDS AND PHRASES
http://view.byu.edu/
**Online software to search the British National Corpus that allows users to search the BNC in various ways; includes tutorials.
●BNC (sample)
http://sara.natcorp.ox.ac.uk/lookup.html
**Online search of the BNC, which provides up to 50 hits.
●Explore N-Grams from the BNC
http://pie.usna.edu/explore.html
**An advanced n-gram search engine for the British National Corpus.
●Daily Newspaper Corpora
http://glossa.fltr.ucl.ac.be/scripts/gtoday/gtoday.pl
**An online KWIC concordancer that searches one or more newspapers in different languages.
●Business Letter Corpus Online KWIC Concordancer
http://ysomeya.hp.infoseek.co.jp/
**An online KWIC concordancer that searches a choice of business letters, personal letters, letters of certain famous people, and a few literary works.
●Corpus Access
http://clwww.essex.ac.uk/cgi-bin/w3c/w3c
**A KWIC program that also produces an output of sentences or paragraphs.
●Web Concordancer (English)
http://www.edict.com.hk/concordance/WWWConcappE.htm
**A web-based concordancer which can be used to search corpora that are provided, such as works of literature, student writing, and newspapers.
●Search PICLE and comparable corpora
http://elex.amu.edu.pl/~przemka/concord2advr/search_adv_new.html
**Concordance that searches a choice of text, including Polish EFL students essays, UK academic textbooks, etc.
●WebCONC
http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi?sprache=en&art=google
**A KIWK concordancing program that allows users to the web pages, or certain web pages, as a concordance. Options include how many pages to search, languages, the number of characters before and after, and which web pages to search.
●Corpus Search (KWIC concordance)
http://www.tooyoo.l.u-tokyo.ac.jp/~kmatsum/corpus/myconc.html
**For English, Estonian, Finnish, Mari, and Turkish.
●Middle English Verse and Prose
http://www.hti.umich.edu/c/cme/
**Various types of searches (simple, Boolean, proximity, etc.) for a Middle English corpus.
●Great Books Online: Bartleby.com/
http://www.bartleby.com/
**A variety of fiction and nonfiction books and references online, including dictionaries, thesauruses, and encyclopedias.
●Michigan Early Modern English Materials
http://www.hti.umich.edu/m/mememl/
**Various types of searches, including basic, Boolean, and proximity searches.
● Michigan Corpus of Academic Spoken English (MICASE)

http://micase.umdl.umich.edu/m/micase/
**A browsable or searchable corpus of nearly two million words made up of transcripts of academic speech events.
●The Middle English Collection
http://etext.lib.virginia.edu/collections/languages/english/mideng.browse.html
**A searchable collection of Middle English texts (some accessible only to users from the University of Virginia).
●Online concordancers
http://132.208.224.131/concordancers/
**Annotated links to four concordances in English and French.
●Penn Treebank Online
http://www.ldc.upenn.edu/ldc/online/treebank/
**Searchable corpora with and without part of speech tags.
●HTI Public Domain Modern English Collection
http://www.hti.umich.edu/p/pd-modeng/
**A resource compiled by the University of Michigan using mainly literature on the Internet.
●Virtual Language Centre Web Concordancer.
http://vlc.polyu.edu.hk/concordance/
**A concordancer that allows users to search English, Chinese, French, Japanese, or parallel texts.
●Web Concordancer
http://www.edict.com.hk/concordance/
**A concordancer that allows users to search English, Chinese, French, Japanese, or parallel texts.
●Web Concordancer (English)
http://vlc.polyu.edu.hk/concordance/WWWConcappE.htm
**A concordancer that allows users to search English using various corpora, including the Brown Corpus and the LOB, as well as texts such as the Sherlock Holmes stories and student writing.
●The Web Concordances and Workbooks
http://www.dundee.ac.uk/english/wics/wics.htm
**Concordances of some classic poetry, some with workbooks to help users.
●WebCorp
http://www.webcorp.org.uk/
**A concordancer which makes use of Internet search engines.
●Concordancing Software
http://www.nsknet.or.jp/~peterr-s/concordancing/specs.html
**A web site with links to various types of software that can be used to make concordances, with information about each. Some are free.
●WordSmith Tools
http://www.lexically.net/wordsmith/
**Downloadable premium lexical analysis software.
●Using WordSmith to Analyse Health Texts
http://www.ling.lancs.ac.uk/staff/paulb/206/health.htm
**Explanation of how to use WordSmith.

(Site in Japanese)

●WebGrep for NESS 6800
http://cow.lang.nagoya-u.ac.jp/program/webgrep/webgrepNESS.html
**杉浦研究室が収集した英語学習用英日対訳例文集。6800 の例文がある。
●Shogakukan Corpus Network
http://www.corpora.jp/
**BNC を日本語のインターフェースで利用できる。有料
●WordbanksOnline
http://www.corpora.jp/~scn/wordbanks.html
**Bank of English を日本語のインターフェースで利用できる。有料
●TXTANA
http://www.biwa.ne.jp/~aka-san/tour_overview.htm
**TXTANA の解説
●Wordsmith を使ってみよう
http://www11.ocn.ne.jp/~iskwshin/wordsmith.html
**Wordsmith の使用方法の解説
●Concordance
http://www.concordancesoftware.co.uk/
**A concordancer that can be used free for 30 days.
●小学館コーパスネットワーク
http://www.corpora.jp/
**BNC の有料検索が日本語のインターフェイスで可能

4.8 Tagging

●Corpus Linguistics
http://www.athel.com/corpus.html#Taggers
**Briefly annotated list of links to tagging programs.
●ACOPOST
http://sourceforge.net/projects/acopost/
**Downloadable part-of-speech tagger.
●Apple Pie Parser
http://nlp.cs.nyu.edu/app/
**A parsing program available on a web page or by ftp.
●CLAWS part-of-speech tagger for English
http://www.comp.lancs.ac.uk/ucrel/claws/

**Information about a part-of-speech tagger for annotating corpora developed at Lancaster University, with links to academic papers.

●Free CLAWS WWW trial service
http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/trial.html
**Sample trial of the CLAWS program.

●Eric Brill's Tagger
http://www.cs.jhu.edu/~brill/
**Link to a tagging program.

●OAK System
http://nlp.cs.nyu.edu/oak/
**An English analyzer that includes “a sentence spliter, a tokenizer, a POStagger, a stemmer, a chunker, a Naned Entity (NE) tagger, a dependency analyzer, a parser, a function tagger and a regularizer.” Currently in development, but interested users can email the researcher and request the software.

●SVMTool
http://www.lsi.upc.es/~nlp/SVMTool/
**Information about software for part-of-speech tagging. Links to online demo and download.

●The Stanford Natural Language Processing Group
http://nlp.stanford.edu/software/tagger.shtml
**A downloadable part-of-speech tagger.

●Statistical natural language processing and corpus-based computational linguistics: An annotated list of Resources
http://www-nlp.stanford.edu/links/statnlp.html
**A large and extensively annotated list of links to resources, including tools, corpora, and literature.

●Syntatic tree tagged corpus
http://kibs.kaist.ac.kr/beginner/kbase3-e.htm
**An explanation for a tagger for Korean which expresses each sentence as a syntactic tree.

●TnT -- Statistical Part-of-Speech Tagging
http://www.coli.uni-saarland.de/~thorsten/tnt/
**A part-of-speech tagger that is free but subject to a license agreement. Can be trained on different languages and tag sets.

●TreeTagger - a language independent part-of-speech tagger
http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
**A downloadable part-of-speech tagger for English, German, Italian, Spanish, French and old French.

●WebTagger: Brill's Tagger を使った英語品詞タグ付与
http://prairie.lang.nagoya-u.ac.jp/program/dobrill.html
**A web-based tagger where users can enter text to be tagged.

4.9 Parallel corpora

●CRATER Multilingual Aligned Annotated Corpus
http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html
**Parallel corpora in English, Spanish, and French, with tags.

(Site in Japanese)

●日英対応付けコーパスの検索
http://www.kotonoba.net/~snj/cgi-bin/text-search/text-search.cgi
**120 以上のパラレルコーパスを使用して検索

●日英新聞記事対応付けデータ (JENAAD)
http://www2.nict.go.jp/jt/a132/members/mutiyama/jea/index-ja.html
**読売新聞と英文読売の対訳コーパス

●パラレルコーパス検索
http://www.eng.ritsumei.ac.jp/asao/corpus/ej.html
**朝尾先生のパラレルコーパス

4.10 Learner corpora

●Corpus of English by Japanese Learners
http://www.eng.ritsumei.ac.jp/lcorpus/
**A corpus of Japanese learners’ English, based on tasks and topics.

●Learner Corpus around the World
http://leo.meikai.ac.jp/~tono/lcorpuslist.html
**Links to corpora from learners of English from different countries.

●Learner Corpus: Resources by Tono
http://leo.meikai.ac.jp/~tono/lcresource.html
**Resources related to learner corpora.

●Search PICLE corpus
http://elex.amu.edu.pl/~przemka/PICLE_search.php
**Links to various resources.

●Advanced Writing in EFL
http://www.geocities.com/writing_site/thesis/
**A PhD based on corpus-based research.

●Cambridge Learner Corpus
http://www.cambridge.org/elt/corpus/clc.htm
**A corpus of learner English which can only be used by users associated with CUP.

●THE LONGMAN LEARNERS' CORPUS
http://www.longman-elt.com/dictionaries/corpus/lclearn.html
**Information about the Longman Learners’ Corpus.

●JEFLL Corpus

http://leo.meikai.ac.jp/~tono/jefll.html
**Description of the Japanese EFL Learner Corpus, its purpose, annotation, etc..
●LINDSEI: Spoken Learner Language: the Lindsei Database
http://cecl.fltr.ucl.ac.be/research%20learner%20corpora.html#lindsei
Description, with contact information, of the **Louvain **In**ternational **D**atabase of **S**poken **E**nglish **I**nterlanguage.
●PELCRA
http://pelcra.ia.uni.lodz.pl/intro_en.php
**Information about and link to the Polish and English Language Corpora for Research and Applications.

(Site in Japanese)

●学習者コーパスデータ
http://www.tomoko-kaneko.com/corpus.html
**研究用にデータを入手できる
●国際学習者コーパス ICLE/LINDSEI 最新の動向
http://www.tomoko-kaneko.com/icle.html
**詳しい情報
●日本人 1200 人の英語スピーキングコーパス
http://www.alc.co.jp/edusys/refecorpus.htm
**世界最大規模の学習者発話コーパス
NICT JLE コーパス（SST コーパス）
●Z 英語学習者コーパス No. 1（中学校英語）
http://home.hiroshima-u.ac.jp/d052121/eigo1.html
***Hiroshima English Learners' Corpus*
●生徒の語彙力を伸ばすために
http://tb.sanseido.co.jp/english/newcrown/pdf/ten-sp01/05.pdf
**論文

4.11 Training for corpus linguistics

●Concordances and Corpora
http://www.georgetown.edu/faculty/ballc/corpora/tutorial2.html
**A thorough tutorial, including information on what corpora are with examples of different types, definitions of various corpus-related terms, etc.
●Concordancing
http://www.ecml.at/projects/voll/our_resources/graz_2002/ddrivenlrning/concordancing/concordancing.htm#Practice
**Online tutorials for concordancing.
●EALC 222: Seminar in Corpus Linguistics Winter 2002
http://www.humnet.ucla.edu/alc/chinese/classes/asian222/
**Course description for seminar on corpus linguistics, including bibliography.
●Corpus Linguistics
http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/contents.htm
**A supplement to *Corpus Linguistics* by Tony McEnery and Andrew Wilson. Includes information about what corpus linguistics is, its history, how it is done, and how it is used in the study of language.
●Corpus Linguistics
http://www.sal.tohoku.ac.jp/ling/corpus1/
**A supplement to *Corpus Lingustics* by Tony McEnery and Andrew Wilson. Includes information about what corpus lingustics is, its history, how it is done, and how it is studied quantitatively.
●Corpus Linguistics
http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/introduction.html
**Resource with information about corpus linguistics, with briefly annotated links.
●Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching
http://www.ctribble.co.uk/text/Palc.htm
**A paper presented at the First international conference: Practical Applications in Language Corpora (1997), University of Lodz, Poland
●A Ten-step Introduction to Concordancing through the Collins Cobuild Corpus Concordance Sampler
http://web.quick.cz/jaedth/CCS-3.htm
**An introduction to how to build a corpus.
●Tutorial: Concordances by Catherine N. Ball
http://www.georgetown.edu/faculty/ballc/corpora/tutorial.html
**Information about using concordancing program.
●ICT4LT Module 2.4 Using concordance programs in the modern foreign languages classroom
http://www.ict4lt.org/en/en_mod2-4.htm
**A module on using corpora and concordancing programs in foreign language classrooms.
●ICT4LT Module 3.4 Corpus linguistics
http://www.ict4lt.org/en/en_mod3-4.htm
**A module introducing the student to corpus linguistics.
●W3-Corpora
http://www.essex.ac.uk/linguistics/clmt/w3c/help/intro/start_page.html
**Information about a corpora search engine, with hints about how to work with corpora.

●Corpus Linguistics
http://www.engl.polyu.edu.hk/corpuslinguist/corpus.htm
**Basic information about corpus linguistics, including what a corpus is, examples of corpora and studies, etc.
●Chi Square Tutorial
http://www.georgetown.edu/faculty/ballc/webtools/web_chi_tut.html
**Explanation of how to use chi squire.
●Web Chi Square Calculator
http://www.georgetown.edu/faculty/ballc/webtools/web_chi.html
**A chi square calculator.
●A few links related to Statistics Education
http://noppa5.pc.helsinki.fi/links.html#stt
**Briefly annotated list of links to information on statistics.
●Free Statistical Software
http://members.aol.com/johnp71/javasta2.html
**Briefly annotated list of links to statistical software.
●Log-likelihood
http://ucrel.lancs.ac.uk/llwizard.html
**A log-likelihood calculator, with an explanation.
●Research Issues in Applied Linguistics
http://www.sal.tohoku.ac.jp/ling/corpus1/index.htm
**A general introduction to corpus linguistics

（Sites in Japanese）

●Ant Conc を使ってみよう
http://www11.ocn.ne.jp/~iskwshin/antconc.html
**石川慎一郎先生による詳しい解説
●インターネットの有効利用 -コーパス言語学の立場から-
http://english.chs.nihon-u.ac.jp/jaecs/workshop/workshop1999/index.html
**この講習会は初心者を対象とした、コーパスに関連した、インターネットの利用法、主要なコーパスを使った検索の実習
●コンピュータ利用による英語教育・英語研究
http://english.chs.nihon-u.ac.jp/jaecs/workshop/workshop2000/index.html
**第１部ではインターネットを使って英語教育・研究に有用な電子テキスト・検索ソフトのダウンロードの実習、第２部では検索ソフトを利用して電子テキストの検索をし、その結果をデータベース化する。
●創ろう！マイデータベース
http://www.babel.co.jp/mtsg/corpus/
**コーパス作成の解説
●ワークショップ 初めてのコーパス検索：WordSmith Tools Version 3 を使って
http://lexis.ias.tokushima-u.ac.jp/wsmith/menu.html
**WordSmith を使用した検索の解説
●WordSmith Tutorials for General Users
http://leo.meikai.ac.jp/~tono/wsmith/index_g.html
**投野先生の WordSmith の活用の解説
●Wordsmith を使ってみよう
http://www11.ocn.ne.jp/~iskwshin/wordsmith.html
**石川慎一郎先生による詳しい解説
●コーパス言語学とは？
http://www.daito.ac.jp/~yamazaki/2001corpus.html
**山﨑ゼミの英語コーパス言語学の概要の解説
●コーパス言語学入門
http://www.tufs.ac.jp/ts/personal/motizuki/lecture/cp2k4/2004_1-j.html
**望月先生の 2004 年のクラス
●《初めてのコンピュータコーパス》
http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/workshop/workshop1997/index.html
**KWIC Concordance for Windows を例に、コンピュータコーパスと、それを処理する検索プログラムについて、初心者を対象に初歩的な実習
●正規表現について
http://infosys.gsid.nagoya-u.ac.jp/~ohna/perl_lesson/regexp.html
**正規表現の解説
●正規表現によるテキスト検索
http://infosys.gsid.nagoya-u.ac.jp/~ohna/re/index.html
**2 日間のワークショップによる解説
●コーパス言語学
http://lexis.ias.tokushima-u.ac.jp/corpus_ling/menu.html
**英語コーパスの言語学の解説
●コーパス言語学のための perl 入門
http://www.eonet.ne.jp/~lago/cat/WorkshopPerl.pdf
**赤瀬川先生の解説
●Perl によるテキスト処理入門
http://infosys.gsid.nagoya-u.ac.jp/~ohna/perl_lesson/index.html
**大名先生の 22 課の詳しい解説
●Google による英文用例検索― 全文検索型サーチエンジンを利用した用例検索 ―
http://infosys.gsid.nagoya-u.ac.jp/~ohna/ggl/
**大名先生の詳しい解説
●「おこしやす」音声を書き起こすソフト
http://www12.plala.or.jp/mojo/
**MP3、WAVE などの音声を聞いて、書き起こすのに適したソフト、スピードのコントロールや割り当てたキーでストップとスタートが可能（無料）

●学生の作品
http://www.daito.ac.jp/~yamazaki/99ronbun.htm
**山﨑先生のゼミの学生の作品
●BlackBox
http://aoki2.si.gunma-u.ac.jp/BlackBox/BlackBox.html
**使用者のデータを統計解析する便利なサイト
●納得！エクセル統計分析
http://w3.cc.nagasaki-u.ac.jp/contrib/Excel/excel1.html
**多少難解
●種々の統計の計算が WEB 上でできる。
http://aoki2.si.gunma-u.ac.jp/calculator/
**27 項目の計算が可能
●統計処理と構造解析
http://oscar.lang.nagoya-u.ac.jp/tech/stat/
**統計の分りやすい解説
●統計学自習ノート
http://aoki2.si.gunma-u.ac.jp/lecture/index.html
**初心者向けの統計の情報
●WWW で統計を学習しよう
http://www.ec.kagawa-u.ac.jp/%7Ehori/statedu.html
**統計に関する膨大な資料集

4.12 Analyzing language using corpora

●Open Text Summarizer
http://libots.sourceforge.net/
**Text summarizing software, which decides which sentences are important in a passage and uses them to make a summary.
●SweSum - Automatic Text Summarizer by Martin Hassel and Hercules Dalianis
http://swesum.nada.kth.se/index-eng-adv.html
**The user enters a text (or a URL of a text to be summarized), keywords, the type of text, the language of the text, what percentage of the text should be used in the summary, etc., and the program produces a summary.
●The Word Sketch Engine
http://www.sketchengine.co.uk/
**"The Word Sketch Engine (WSE, also known as the Sketch Engine) is a new Corpus Query System incorporating word sketches, grammatical relations, and a distributional thesaurus. A word sketch is a one-page, automatic, corpus-derived summary of a word's grammatical and collocational behaviour."

(Sites in Japanese)

●字、単語、文、パラグラフの数：読みやすさ
http://www.cis.doshisha.ac.jp/kkitao/Japanese/library/resource/corpus/word.doc
**Word による解析の解説
●語彙頻度と文に分ける
http://oscar.gsid.nagoya-u.ac.jp/program/perl/mlu2.html
**Software that provides the word frequency, the number of sentences, the number of words, and the words per sentence.
●語彙頻度と文に分ける(解説)
http://www.cis.doshisha.ac.jp/kkitao/Japanese/library/resource/corpus/mlu2[1].htm
**杉浦先生のプログラムの解説
●語彙の頻度表
http://www.cis.doshisha.ac.jp/kkitao/Japanese/library/resource/corpus/indexer.doc
**Catherine N. Ball の Web Frequency Indexer の使用方法の解説
●文毎に行を変える
http://www.cis.doshisha.ac.jp/kkitao/Japanese/library/resource/corpus/hidemaru.doc
**秀丸の正規表現の解説
●Text-Based Concordances (v. 2)
http://www.cis.doshisha.ac.jp/kkitao/Japanese/library/resource/corpus/text.htm
**手持ちのテキストのすべての語がどのように文章で現われるかを KWIC 形式でアルファベット順にリストします。
●N-Gram Phrase Extractor (v.3)
http://www.cis.doshisha.ac.jp/kkitao/Japanese/library/resource/corpus/n-gram.htm
**手持ちのテキストのすべての N 語の組み合わせがどのように文章で現われるかを KWIC 形式でアルファベット順にリストします。
●KWIC,右と左のソート
http://uluru.lang.osaka-u.ac.jp/~k-goto/web_kwic.html
**後藤氏の作で、フランス語やドイツ語にも対応
●JACET8000 語による語彙の頻度による分析
http://www.cis.doshisha.ac.jp/kkitao/Japanese/library/resource/corpus/j8lm.cgi.htm
**「JACET 8000 語のチェック」の使用方法の解説と例
●「JACET 8000 語のチェック」の使用方法の解説と例の結果を頻度グループに分ける解説
http://www.cis.doshisha.ac.jp/kkitao/Japanese/library/resource/corpus/j8lm.cgi.doc
**JACET8000 語の表示
●単語の頻度数検索と KWIC 表示
http://www.babel.co.jp/mtsg/corpus/kwic/kwic.htm
**KWIC Concordance for Windows (塚本　聡) の解説

●スラッシュ・リーディング支援システム
http://lengua.cc.kyushu-u.ac.jp/english/sr/
**田中省作氏作で、箱に英語の文章を入れれば、切れ目にスラッシュを入れて表示する。初級、中級、上級を選べる
●コロケーション
http://www.slacorpus.com/programs/jpn.html
**永野友雅氏の製作で、5600 万語のコーパスを利用して、２語のコロケーションを検索する。学習者の２語のコロケーションで、よりよいものを表示するものもある。

perl のプログラム
●トークンをレンマに (朝尾先生の公開)
http://www.eng.ritsumei.ac.jp/asao/software/lemma/
**テキストに現れるトークンをそのままレンマに置き換えたい場合がある。lemma.pl はこのような目的に使える。
●英語と日本語の混じった文章から日本語を削除する。
http://www.cis.doshisha.ac.jp/kkitao/Japanese/library/resource/corpus/del_2byte.doc
**複数のテキストを同時に処理可能

4.13 Electronic texts (non-copyrighted corpora)
●4Literature
http://www.4literature.net/
**A collection of more than 2000 books, stories, poems, and documents, which can be searched or browsed by title or author.
●Alex Catalogue of Electronic Texts
http://www.infomotions.com/alex/
**A collection of documents in the public domain from American and English literature and Western philosophy, which can be searched or browsed by title or author.
●Alex Catalogue of Electronic Texts
http://www.infomotions.com/alex2/
**A full-text indexed collection of classic American and English literature as well as Western philosophy in the public domain and written or translated into English.
●The ARTFL Project
http://humanities.uchicago.edu/orgs/ARTFL/
**Literature and reference materials in French and English.
●Bartleby.com: Great Books Online
http://www.bartleby.com/sv/welcome.html
**A searchable or browsable collection of reference works, verse, fiction, and non-fiction.
●biblimania.com
http://www.biblimania.com/contents/complete53.php
**Briefly annotated list of online texts.
●Center for Electronic Texts in the Humanities
http://www.ceth.rutgers.edu/
**Annotated list of links to various online collections of texts, some open and some restricted.
●English corpora
http://www.corpora4learning.net/resources/corpora.html
**Short descriptions of and links to well-known English corpora, including American English, British English, Singapore English, English as a lingua franca, and historical English.
●Eserver
http://eserver.org/
**Briefly annotated list of links to a wide variety of texts.
●The Etext Archive
http://www.etext.org/index.shtml
**Electronic texts related to politics and religion, and e-zines, fiction, and poetry.
●The Etext Center at the University of Virginia Library
http://etext.lib.virginia.edu/
**A collection of e-texts, “including classic British and American fiction, major authors, children's literature, American history, Shakespeare, African-American documents, the Bible, and much more. Available in Microsoft Reader, PalmReader, and soon ".pdf" formats.” Can be browsed or searched.
●The Humanities Text Initiative
http://www.hti.umich.edu/
**A broad collection of freely available e-texts, including versions of the Bible, research papers and reports, and public papers of American Presidents.
●ILT Digital Text Projects
http://www.ilt.columbia.edu/publications/digitext.html
**Writings of various philosophers, including Descartes, Emerson, Kant, and Locke.
●International Corpus of English
http://www.ucl.ac.uk/english-usage/ice/index.htm
**Corpora in various Englishes, available on CD-ROM or as a download.
●Internet Archive
http://www.archive.org/
**Archive of movies, music, and texts. Also has the Wayback Machine, which archives web pages.
●The Internet Classics Archive

http://classics.mit.edu/
**Browsable and searchable texts of classic literature.
●The Library of Congress
http://www.loc.gov/
**The web page of the Library of Congress, which has a wide variety of texts related to American culture and history.
●Linguistic Data Consortium (LDC)
http://www.ldc.upenn.edu/
**A variety of corpora, some tagged or analyzed, in different languages.
●Linguistic Data Resources on the Internet
http://www.sil.org/linguistics/etext.html
**List of links, some briefly annotated, to electronic text centers, digital libraries, text collections, etc.
●List of Corpora
http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/index2.html#speech
**List of links to speech corpora, some briefly annotated.
●List of Corpora
http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/
**List of links to various corpora, mostly unannotated.
●List of E-text Archives
http://lang.nagoya-u.ac.jp/~matsuoka/e-texts.html
**Unannotated list of etext collections.
●Mike Nelson's Business English Lexis Site
http://users.utu.fi/micnel/business_english_lexis_site.htm
**Business English words, based on corpora research for Nelson's MA thesis. Includes links to the thesis and other business English-related materials.
●OLAC, the Open Language Archives Community
http://www.language-archives.org/
**An archive developed by an organization whose purpose is "creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources."
●Online Books
http://home.eplus-online.de/veittes/online.html
**Unannotated list of links to online book collections.
●The Online Books Page
http://digital.library.upenn.edu/books/
**Very extensive lists of links to online books, some annotated, browsable by title and author. Also has such special features as information about "banned books" and a celebration of women writers.
●The Oxford Text Archive
http://ota.ahds.ac.uk/
**Browsable and searchable collection of texts, which themselves are searchable.
●Project BookRead
http://tanaya.net/BookRead/
**A large online collection of books, browsable by author or title.
●Project Gutenberg
http://www.gutenberg.org/
**An online catalogue of about 17,000 books in a wide variety of languages.
●SunSITE Digital Collections
http://sunsite.berkeley.edu/Collections/
**Collection of etexts, including texts on California history, online reading for classes, and government documents.
●Words' Realm
http://www.heise.de/ix/raven/Literature/Etext.html
**Mainly unannotated list of links to etexts.
●KidPub
http://www.kidpub.org/kidpub/
**Web page where children can post stories they have written, read other children's stories, discuss stories, etc.. (copyrighted)
●American Memory Home
http://memory.loc.gov/ammem/
**Collections of documents and other writings related to various aspects of American history and culture.
●The Berkeley Digital Library
http://sunsite.berkeley.edu/
**An archives related to American culture and history, especially related to California.
●Women Writers Project
http://www.wwp.brown.edu/
**A database of early modern writing by women.

Mass Media

●ABC News
http://abcnews.go.com/
**Web page of ABC news, with news stories, some news story videos, etc.
●CNN Transcripts
http://transcripts.cnn.com/TRANSCRIPTS/
**Transcripts of CNN programs, categorized by date and program.
●eNewswires.com
http://www.enewswires.com/

**Links to online news sites and news stories.
●The Internet Public Library: Newspapers
http://www.ipl.org/div/news/
**List of links to newspapers all over the world.
●11 Mass Media (News Sources)
http://www.cis.doshisha.ac.jp/kkitao/online/www/referenc.htm#mass
**Links to mass media web pages, some with annotations.
●News & Periodical Resources on the Web
http://www.loc.gov/rr/news/lists.html
**Lists of links to news web sites, some with annotations.
●News Directory
http://www.newsdirectory.com/
**Unannotated list of links to mass media web pages, arranged by type and region.
●Pathfinder
http://www.pathfinder.com/pathfinder/index.html
**Web page for 18 magazines, including *Time, People, Fortune, and EW.*
●U.S. News & World Report
http://www.usnews.com/usnews/home.htm
**Web page for *US News and World Report*, with current news and feature articles as well as information about colleges and graduate schools, etc.
●World newspapers online
http://www.actualidad.com/
**List of links to news sources, with information.

Movies

●Drew's Script-O-Rama
http://www.script-o-rama.com/oldindex.shtml
**A very large collection of movie and TV scripts, as well as other film- and tv-related resources.
●Find -A-Script
http://www.angelfire.com/film/thegreenlightzone/findscript.html
**A large collection of movie scripts, with links to other web pages where scripts are available.
●Movies
http://www.cis.doshisha.ac.jp/kkitao/online/www/teacher.htm#movie
**Movie-related web pages, some with annotations.
●Movie Scripts and Screenplays Web Ring Home Site
http://www.moviescriptsandscreenplays.com/
**Links to web sites where movie and tv scripts are available.
●Search the Internet Movie Database
http://us.imdb.com/search
**The search page for the IMDb, where users can search by movie title, cast/crew name, character name, etc.
●Scripts A-M
http://www.movie-page.com/movie_scripts.htm
**Scripts for a large number of movies, plus a forum for discussion.
●Simply Scripts
http://www.simplyscripts.com/
**Web site with movie scripts, tv scripts, etc.

Speeches

●American Rhetoric: Top 100 Speeches by Rank
http://www.americanrhetoric.com/top100speechesall.html
**A list of 100 speeches, some of which users can listen to.
●Speech Corpora
http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/index2.html#speech
**Annotated list of links to speech corpora.
●Online Speech Bank
http://www.americanrhetoric.com/speechbank.htm
**A collection of audio and video versions of American speeches, sermons, legal proceedings, etc.
●Inaugural Addresses of the Presidents of the United States
http://www.bartleby.com/124/
**Manuscripts of Americans Presidents’ inaugural addresses, with brief explanations of the context.

spoken corpora

●TRAINS spoken dialogue transcriptions
http://www.cs.rochester.edu/research/speech/93dialogs/
**A corpus of short dialogues (transcribed online and available on CD ROM), up to 13 minutes long, but mostly less than 5 minutes. Each dialogue has a problem to solve.

5. Conclusion

As readers can see from browsing “Resources for Corpus Linguistics,” there are a great many resources related to corpus linguistics available on the Internet. Many of these are valuable even for users who are not professional researchers in corpus linguistics. Teachers of English and students of English can make use of these resources to better understand specific points of

the English language, to make materials, and so on.

The author hopes that both those who have had experience with corpus linguistics and those who have a potential interest in the subject will find useful resources here.

Acknowledgement

The author would like to acknowledge Dr. S. Kathleen Kitao, who helped to make annotations for the links.

References

Aarts, J. & Meijs, W. (Eds.) (1984). *Corpus Linguistics*. Amsterdam: Rodopi.

Francis, W. N. & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Johansson, S. & Hofland, K. (1989). *Frequency analysis of English vocabulary and grammar based on the LOB Corpus*. Oxford: Clarendon Press.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijme,. & B. Altenberg, (Eds.). *English corpus linguistics: Studies in honour of Jan Svartvik*. London: Longman, 8-29.