コーパス紹介

2016年度　春学期

水曜2限：言語教育学演習

目次

1. General Corpus 3
   1.1 British National Corpus (BNC) 4
   1.2 Corpus of Contemporary American English (COCA) 8
2. Spoken Corpus 12
   2.1 London-Lund Corpus of Spoken English (LLC) 13
3. ESP Corpus 15
   3.1 British American Spoken Corpus (BASE) 16
   3.2 British American Written Corpus (BAWE) 18
   3.3 Michigan Corpus of Academic Spoken English (MICASE) 20
   3.4 Professional English Research Consortium (PERC) 23
4. Diachronic Corpus 26
   4.1 Helsinki Corpus 27
5. Learner Corpus 30
   5.1 Cambridge Learner Corpus 31
   5.2 International Corpus of Learner English (ICLE) 34
   5.3 NICT JLE Corpus 38
6. Web as Corpus 41
   6.1 enTenTen 42
   6.2 Web as Corpus 49

1. GENERAL CORPUS

1. GENERAL CORPUS

1.1 British National Corpus

BNC

The British National Corpus

- 書きことば、話し言葉をあわせて1億語から成る、世界最大のイギリス英語コーパスである。
- 1億語の内訳は約90%が書き言葉、10%が話し言葉で、1985年から1990年までに出版され、記録され、あるいは話されたテキストが集められている。

BNC

The British National Corpus

- BNCの各テキストには、すべての単語に品詞情報（POS tag)が付与されている。
- その他、文書の構成（見出し、段落など）のみならず、分野、発行年代、書籍名、著者、性別、年齢、発行場所、出版社などの詳細な書誌情報も付与されている。

BNC

The British National Corpus

- アクセス：https://www.sketchengine.co.uk/
- 使用有料のコーパスですが、東京外国大学キャンパスネットワークを利用する場合、自動的にアクセスできる。
- 以上のリンクをクリック→Log in→ Home→ BNCを選択。

1. GENERAL CORPUS

1.2 Corpus of Contemporary American English

COCA

Corpus of Contemporary American English

- オンライン上で無料で利用できるコーパスとして4億語から成る現在最大の均衡コーパスである。BNCの4倍の大きさがある。
- Brigham Young Universityの Mark Davies教授によって作られた。

COCA

Corpus of Contemporary American English

- 現在も更新中であるため、最近の言語変化の研究に最適である。
- 話し言葉、小説、人気雑誌、新聞、学術雑誌の5ジャンルがある。
- http://corpus.byu.eduにアクセスし、“Corpus of Contemporary American English (COCA)” をクリックする。

COCA

Corpus of Contemporary American English

- 利用者登録をしなくても、10回の程度で検索でする。
- コーパスを使い続けたい場合は、登録が必要である。
- 登録が完了したら、自分のユーザでログインする。

2. SPOKEN CORPUS

2. SPOKEN CORPUS

2.1 London-Lund Corpus of Spoken English

LLC:(The London-Lund Corpus of Spoken English)

発案者:Jan Svartvik(Lund University)

編集期間:1959-1990

規模:500,000語

言語:イギリス英語

基金:Bank of Sweden Tercentenary Foundation, L.M Ericsson

アクセス:Oxford Text Archive (http://ota.ahds.ac.uk)

3. ESP CORPUS

3. ESP CORPUS

3.1 British American Spoken Corpus (BASE)

BASE：(British Academic Spoken English)

発案者：Hilary Nesi, Paul Thompson（Universities of Warwick and Reading)

編集期間：2000ー2005

規模：164,4942語

言語：英語

基金：Universities of Warwick and Reading,
BALEAP,EURALEX, The British Academy

アクセス：Oxford Text Archive (http://ota.ahds.ac.uk)

3. ESP CORPUS

3.2 British American Written Corpus (BAWE)

BAWE: (British Academic Written English)

発案者：Nesi Hilary, Gardner ,Sheena, Thompson

編集期間：2004-2007

言語：英語

基金：Economic and Social Research Council

アクセス：Oxford Text Archive (http://ota.ahds.ac.uk)

3. ESP CORPUS

3.3 Michigan Corpus of Academic Spoken English (MICASE)

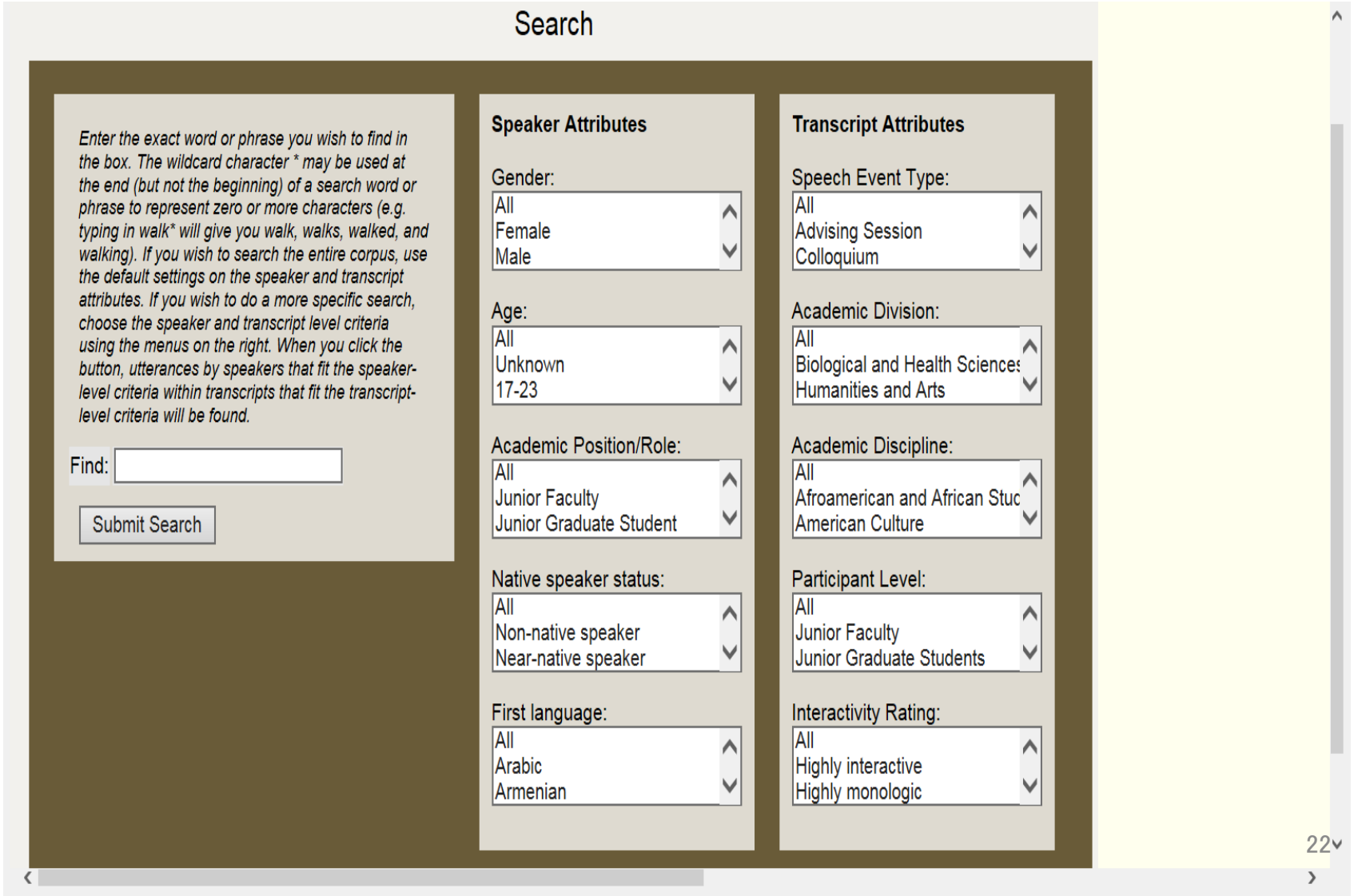
MICASE

(http://quod.lib.umich.edu/cgi/c/corpus/corpus?page=home;c=micase;cc=micase)

- Michigan Corpus of Academic Spoken English(MICASE)
- Specialized corpus
- 1997年 制作開始、2002年 公開
- 約180万語のスピーチデータ
- 講義、ディスカッション、セミナーなど
- 登録の必要はなく、ウェブサイトから自由に利用可能
- 検索画面から話者とトランスクリプトの属性が指定できる

検索画面

(http://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple)



3. ESP CORPUS

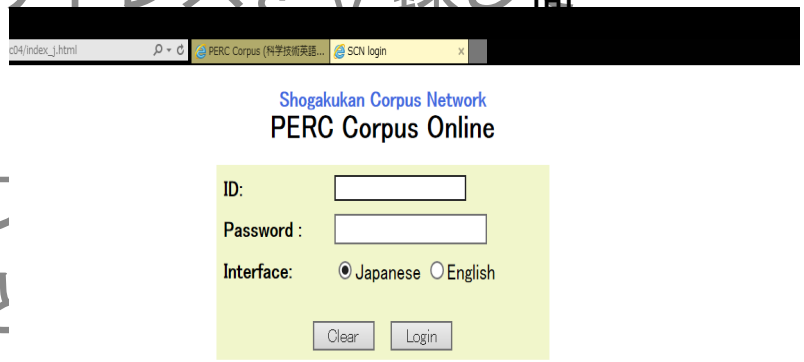
3.4 Professional English Research Consortium (PERC)

PERC Corpus

(https://scn.jkn21.com/~percinfo/index_j.html)

- Professional English Research Consortium (PERC)
- Specialized corpus
- 2001年 制作開始
- 使用するには名前、メールアドレスを登録してIDとパスワードを得る
- 約1700万語の学術雑誌論文
- 医学、生物、物理、数学など22の領域

ログイン画面



検索画面

・検索する前に必ずお読みください。

・語(句)検索の簡単な使い方と注意

4. DIACHRONIC CORPUS

4. DIACHRONIC CORPUS

4.1 Helsinki Corpus

Helsinki Corpus

＜URL＞http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html

・最新版がもう少しで完成する（Oxford Text ArchiveとVARIENG unit)から利用可能になる予定

・初版であるHelsinki Corpus TEI XML Edition 2011はHelsinki Corpus of English Texts (1991)に基づいて作成された

Helsinki Corpus

- 8世紀から18世紀までの1000年間の英語テキストを含んでいる
- 2千から2万語の様々な400以上の文書から150万語を収集
- 3つのセクションに分けられる（Old English,Middle English, and Early Modern English）・それぞれのセクションは3，4の70から100年単位のさらに小さい区分に分けられる

それぞれのテキストサンプルには「日付」、「方言」、「ジャンル」、「その他の特徴」を表すparameter codesがある

- 文法や統語的なコードを含んでいない（初期に公開されたバージョン）
- コーパスの代表性に関わる要素としては①地域（例）方言②社会的差異（例著者の性別、年齢、社会や教育的背景③ジャンル（例）どのようなタイプの書き言葉か、目的など
- ①方言はOld EnglishとMiddle Englishのセクション、②社会的差異はLate MiddleとEarly Modern English ③ジャンルは年代が後の方が内容が充実している

5. LEARNER CORPUS

5. LEARNER CORPUS

5.1 Cambridge Learner Corpus

Cambridge LC　webからの情報

・Cambridge English Corpus の一部である。

・イギリス各地から集めた話し言葉の録音を非常にたくさん集めている

・現在も新たな情報を集めるために参加者を募集している

・英語学習者の典型的なミスを識別することができる

Cambridge LC　検索で得られる内容

・トピックやレベルにあった最も適切な単語、フレーズを選択できる

・スピーキング力やリスニング力を鍛えるために、話し言葉を分析している

5. LEARNER CORPUS

5.2 International Corpus of Learner English

ICLE

＜URL＞http://www.uclouvain.be/en-cecl-icle.html

・様々な母語(Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana, Turkish)の上級英語学習者のエッセイのデータを集めた英語学習者の国際コーパス

・多くの国際提携大学の共同成果

・初版は2002年にCD-ROMで出版された

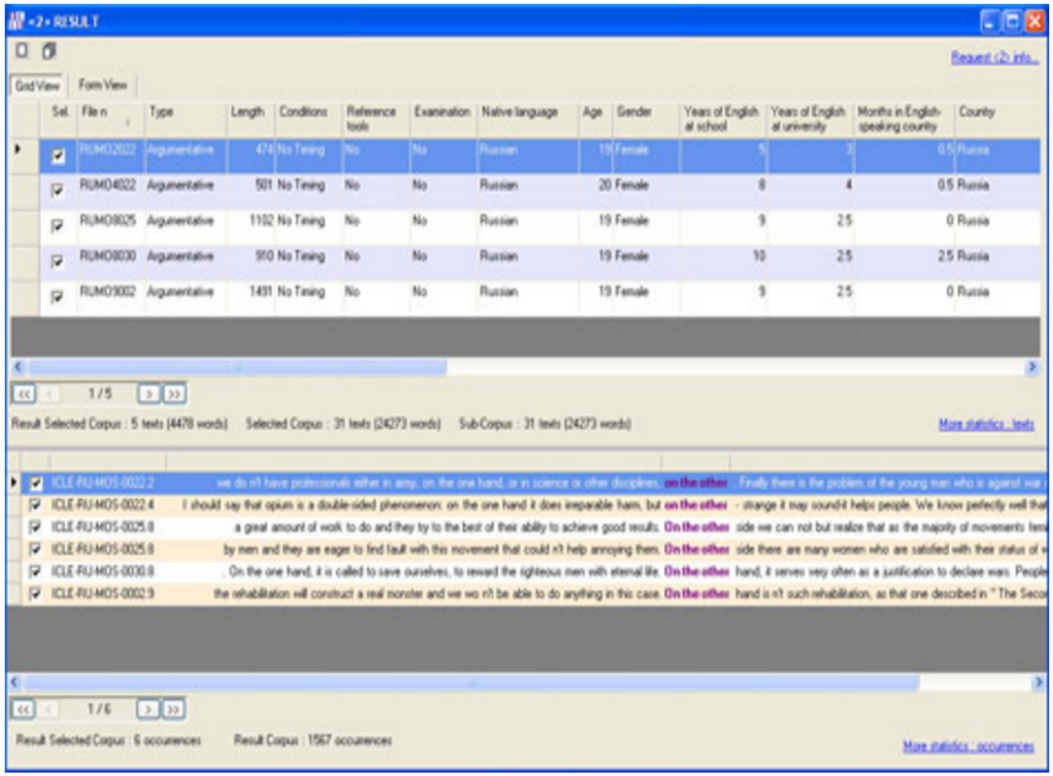
・改訂版であるICLEv2は2009年に出版され用語索引を内蔵したことが特徴になっている

・同じコーパス収集ガイドライン（テスト方法）を用いることで均質的なコーパスになっている

・現在第3版作成中

ICLE 検索手順

1. **built-in concordancer** allowing users to search for word forms, lemmas and/or part-of-speech tags;



語形、レンマ、品詞
タグで
検索可能

Note, however, that the unannotated ICLE texts can be downloaded in txt format and processed with any other software tools (concordancer, POS-tagger, parser, etc.).

ICLE 検索手順

2. **breakdown of the query results** according to the learner profile information:

- Distribution of the occurrences of a linguistic query according to the whole set of ICLE-recorded variables (e.g. native language, gender, age, type of task)

検索結果から学習者の情報を検索可能
（母語、性別、年齢、タスクの種類）

- 'Range' analysis: distribution of ICLE variables for the texts that match a corpus collection or a linguistic query

スクリーンショッ
Dropbox フォ
ました。

5. LEARNER CORPUS

5.3 NICT JLE Corpus

NICT JLE　webからの情報

・日本語を母語とする英語学習者の英語インタビューテストにおける発話を書き起こしたデータを中心に構成されている

・話し言葉の学習者コーパスとしては世界最大級の規模を誇る

・国立研究開発法人情報通信研究機構が株式会社アルク協力のもと構築した

NICT JLE 検索プログラムの方法

主なタグの意味はこのようになっている

- <head version=“1.3”>性別(sex)、海外経験(overseas)、熟達度レベル(SST_Level)等の学習者のプロフィール</head>
- <A>インタビュアーの発話</A>
- <B>学習者（受験者）の発話</B>
- <task>インタビュー内で課せられるタスク部分</task>
- <followup>タスク後の雑談</followup>
- <F>つなぎ言葉（フィラー）</F>
- <R>言い直し</R>
- <SC>言い淀み</SC>

6. WEB AS CORPUS

6. WEB AS CORPUS

6.1 enTenTen

The TenTen corpus family

- “bigger, wider-coverage, cleaner, duplicate-free, and with richer metadata.” (Jakubíček,M. et al.) を目標にした、世界の主要な言語のウェブコーパス（＝ウェブ上の文書を取得し、自動的にデータベース化したコーパス）群。

- アラビア語
- 中国語
- 英語
- フランス語
- ドイツ語
- イタリア語
- 日本語
- 韓国語
- ポルトガル語
- ロシア語
- スペイン語

現在、＋20言語利用可

TenTenコーパス一覧はhttps://www.sketchengine.co.uk/tenten-corpora/参照。

- TenTen = 10^{10}(=100億) から。100億語～規模のコーパス

The TenTen corpus family

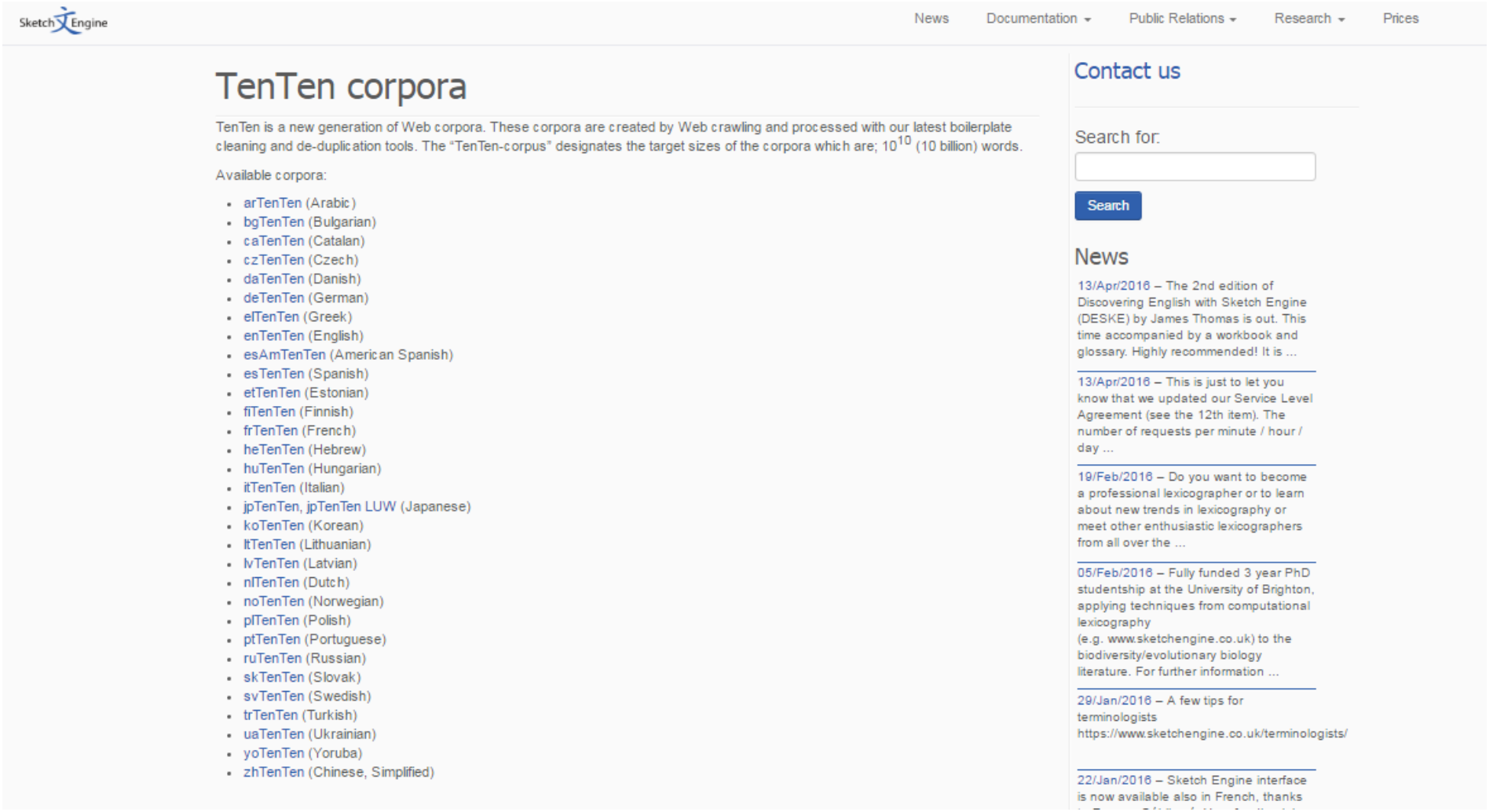
- 2012年～

前進となるいくつかのウェブコーパスをもとに初期の形を構築。
以後１～２年ごとにウェブを再クロール、処理して付け加え

Spiderling ・・・ウェブのあらゆるテキストをダウンロード。
jusText ・・・ウェブページの余計な要素を排除。
Onion ・・・重複を排除。

+Tokenize, lemmatize, part-of-speech tagging

The TenTen corpus family



Sketch Engine

News Documentation Public Relations Research Prices

TenTen corpora

TenTen is a new generation of Web corpora. These corpora are created by Web crawling and processed with our latest boilerplate cleaning and de-duplication tools. The "TenTen-corpus" designates the target sizes of the corpora which are; 10^{10} (10 billion) words.

Available corpora:

- arTenTen (Arabic)
- bgTenTen (Bulgarian)
- caTenTen (Catalan)
- czTenTen (Czech)
- daTenTen (Danish)
- deTenTen (German)
- elTenTen (Greek)
- enTenTen (English)
- esAmTenTen (American Spanish)
- esTenTen (Spanish)
- etTenTen (Estonian)
- fiTenTen (Finnish)
- frTenTen (French)
- heTenTen (Hebrew)
- huTenTen (Hungarian)
- itTenTen (Italian)
- jpTenTen, jpTenTen LUW (Japanese)
- koTenTen (Korean)
- ltTenTen (Lithuanian)
- lvTenTen (Latvian)
- nlTenTen (Dutch)
- noTenTen (Norwegian)
- plTenTen (Polish)
- ptTenTen (Portuguese)
- ruTenTen (Russian)
- skTenTen (Slovak)
- svTenTen (Swedish)
- trTenTen (Turkish)
- uaTenTen (Ukrainian)
- yoTenTen (Yoruba)
- zhTenTen (Chinese, Simplified)

Contact us

Search for:

Search

News

13/Apr/2016 – The 2nd edition of Discovering English with Sketch Engine (DESKE) by James Thomas is out. This time accompanied by a workbook and glossary. Highly recommended! It is ...

13/Apr/2016 – This is just to let you know that we updated our Service Level Agreement (see the 12th item). The number of requests per minute / hour / day ...

19/Feb/2016 – Do you want to become a professional lexicographer or to learn about new trends in lexicography or meet other enthusiastic lexicographers from all over the ...

05/Feb/2016 – Fully funded 3 year PhD studentship at the University of Brighton, applying techniques from computational lexicography (e.g. www.sketchengine.co.uk) to the biodiversity/evolutionary biology literature. For further information ...

29/Jan/2016 – A few tips for terminologists https://www.sketchengine.co.uk/terminologists/

22/Jan/2016 – Sketch Engine interface is now available also in French, thanks

en-TenTen
(TenTen プロジェクトの英語コーパス)

- TenTenプロジェクトに沿って構築された英語のコーパス

（＝ウェブ上の文書を取得し、自動的にデータベース化）

・サイズ：約100億(2012版) ~ 約200億語（2013版)

※Sketch Engine では “English Web 2013（約197億語）” となっている

・地域：アメリカ、イギリスにはタグあり、それ以外はタグなし

・TreeTagger によって lemma や PoSの情報が自動タグ付けされている

en-TenTen コンコーダンス画面

参考

“TenTen corpora.” *Sketch Engine*. 5 June 2016.
<https://www.sketchengine.co.uk/tenten-corpora/>

“enTenTen corpus.” *Sketch Engine*. 5 June 2016.
<https://www.sketchengine.co.uk/enTenTen-corpus/>

Jakubíček,Miloš.,Kilgarriff,Adam.,Kovář,Vojtěch.,Rychlý,Pavel.,Suchomel,Vít. “The TenTen Corpus Family.” at 7th International Corpus Linguistics Conference (2013). Web. 5 June 2016< https://www.sketchengine.co.uk/wp-content/uploads/The_TenTen_Corpus_2013.pdf>

6. WEB AS CORPUS

6.2. Web as Corpus (WAC)

WaC (Web as Corpus) Corpora

- １億語をはるかに超える超大規模コーパス
- ウェブ上の言語資源を自動収集
- 無作為に選ばれた内容語のペアを検索サイトに送り、検索結果のアドレスを「種」としてデータを収集するプロセスを反復することで偏りの少ない言語データを収集する
- Sketch Engineでは、54か国語のWaCが利用可能

さまざまなWaC

- ukWaC
  - 13億語のイギリス英語コーパス
  - Adriano Ferraresiが作成
  - Sketch Engineで利用可能
  - POSタグ付与済み，レマ検索も可能

- JpWaC
  - ４億語の日本語コーパス

ukWaC

Sketch Engine
ukWaC
Prof. Yukio Tono
Home
Search
Word list
Word sketch
Thesaurus
Sketch diff
Corpus info
My jobs
User guide
Save
Make subcorpus
View options
KWIC
Sentence
Sort
Left
Right
Node
References
Shuffle
Sample
Filter
Query look|bring, V.*, up|down 67,777 (43.50 per million)
Page 1 of 3,389 Go Next | Last
site=bbc.c... grammatical function and meaning; as does a colon: look it up ! ...I and always insist on saying "I've
site=bbc.c... knows? In fifty years children could be brought up with this ridiculous accent. It's a shame
site=briti... short caress on his first welcomer, and then looking up , saw one of the shutters partially unclosed
site=clieu... I would recommend them highly if you are looking to tidy up some of the rubbish left behind on your
site=lords... society in which so many people now place the bringing up of children, and particularly their spiritual
site=lords... especially to do all in our power to faithfully bring up the youngsters he has entrusted to us to
site=lords... and family priorities - Malachi 2:14-16 Bringing up youngsters - Ephesians 6:1-4 Questions
site=athel... they reached the crown of the low ridge and looked down at a strip of open veldt, beyond which
site=class... accurate there is no trace of a path. " Looking down from the hilltop, to which cling the ruins
site=bears... smile. Find A Grave Bring out your dead! Or, look them up here and bring 'em out yourself. Via Bored
site=cambr... clinic has recently been re-furbished. Always bring an up to date list of your medication when you
site=sange... characters found anywhere in the name you are looking up . (e.g. A search for "blo" , "log" and "
site=nutco... 2 Orbiter in 1977. This dramatic picture looks down from the Orbiter's viewpoint about 8,000
site=senio... would be getting supper ready for Papa. She looked up at me, and couldn't believe what she saw
site=bbkph... and those in Eastern Europe who started to bring a wall down . Are the conclusions of this book open
site=zetne... embroidery, but ever so often she would look up and smile. Her father was, to all intents
site=zetne... the lounge door opened again. Everybody looked up at the unexpected visitor. Carrying a beautifully
site=zetne... his wrist so she might see better. Then he looked back down and gently touched the crystal. A chipmunk
site=zetne... "She reckoned I was gadget mad, so.." he looked down , sighed, "..so she went to the biggest
site=zetne... he laughed and pointed to the street. She looked up at the snow beginning to fall. He laughed
Page 1 of 3,389 Go Next | Last
115%

参考文献

- 石川慎一郎 2012.『ベーシックコーパス言語学』ひつじ書房.