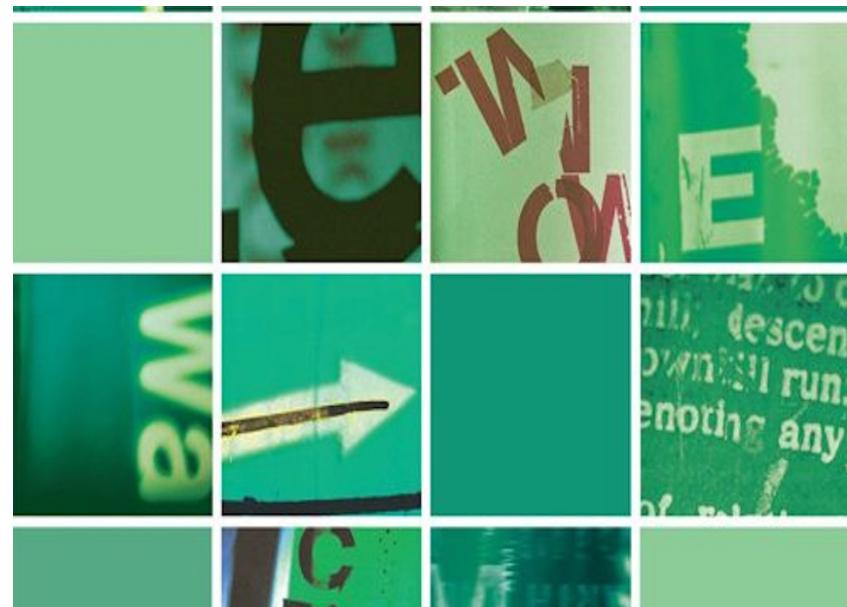


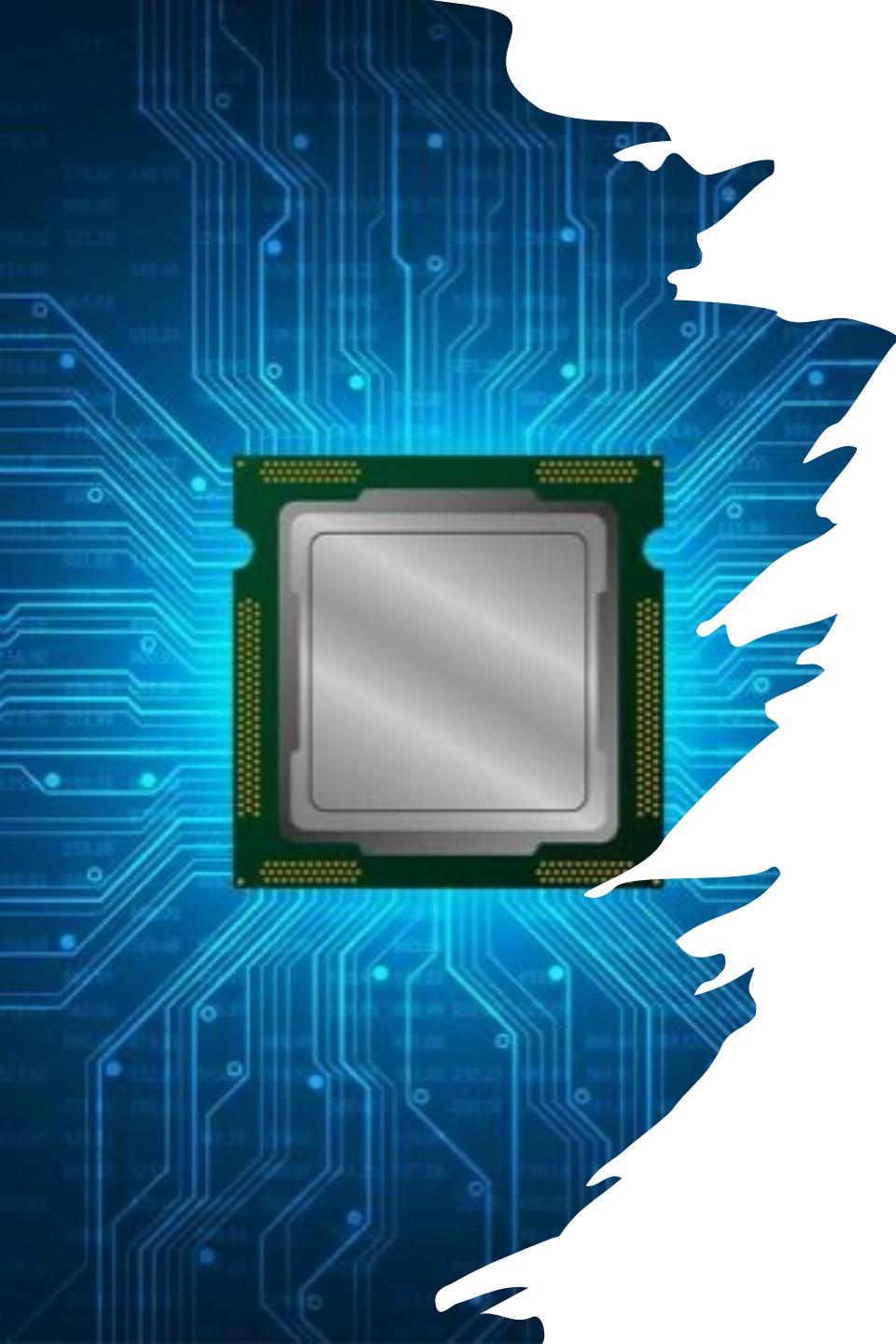


Lecture 7. Corpus linguistics

© Kushneruk Svetlana
Leonidovna

Doctor of Philology, Professor of Chelyabinsk State University





1. The notion of a corpus. Typology of corpus linguistic research

Corpus linguistics revolutionized language studies because it has provided **new ways** of analyzing and describing the use of language

- **Corpus linguistics**
- a **powerful methodology** that can be employed to explore a wide variety of issues related to the use of vocabulary
- **Corpus linguistics**
- an area which focuses upon a **set of procedures**, or **methods**, for studying language



Corpora can be defined as large, principled and ***computer-readable collections of texts*** that allow analysis of patterns of language use across different contexts.

Corpora consist of texts stored in an electronic format, which enables researchers to use special software to **conduct automatic searches** and gain insights into the **structure and regularity of naturally occurring language**.



Important features of corpus-based analysis



it is empirical,
analyzing the actual
patterns of use in
natural texts



it utilizes a large
collection of natural
texts as the basis for
analysis



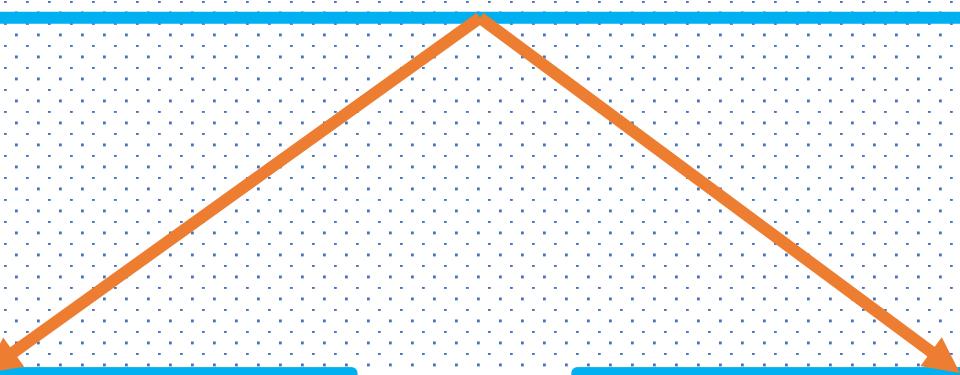
it makes extensive
use of computers for
analysis



it depends on both
quantitative and
qualitative analytical
techniques

Typology of corpus linguistic research.

1.1. Mode of communication



**corpora of written
language**

**corpora of spoken
language**

UK Parliament

[UK Parliament](#) >  > [House of Commons Hansard archives](#) > House of Commons debates by date archive 1988-2016

House of Commons debates by date archive 1988-2016

This is now an archive and the latest debates can be found in [Hansard](#).

Browse House of Commons debates from November 1988 to March 2016 on statements, petitions, oral and written questions and answers. Written Answers from 12 September 2014 are now published on the [Written Questions and Answers](#).

Browse House of Commons debates, statements, petitions, oral and written questions and answers, since November 1988, by date. Alternatively, archived volumes of Commons Hansard are accessible via the right menu.

In this section

[House of Commons Hansard archives](#) 

[House of Commons debates by date archive 1988-2016](#)

[Archived Commons Hansard](#) >

www.publications.parliament.uk/pa/cm/cmhansrd.htm

ICE-GB is the British component of the International Corpus of English (ICE)

Search UCL GO UCL Home » Survey of English Usage » Projects » ICE-GB

Survey of English Usage

- ➊ Home
- ➋ About the Survey
- ➌ Staff
- ➍ Research Projects
 - ➎ The International Corpus of English (ICE)
 - ➏ ICE-GB
 - ➐ DCPSE
 - ➑ Corpus Queries
 - ➒ Next Generation Tools
 - ➓ The English Noun Phrase
 - ➔ The English Verb Phrase
 - ➕ Subordination in Spoken & Written English
 - ➖ Teaching English Grammar in Schools
 - ➗ Research Resources
 - ➘ Software Sales
 - ➙ Mobile Apps
 - ➚ Events

ICE-GB

ICE-GB is the British component of the International Corpus of English (ICE).

ICE began in 1990 with the primary aim of providing material for comparative studies of varieties of English throughout the world.

More than twenty centres around the world are preparing corpora of their own national or regional variety of English. These include

Australia	Malaysia
Cameroon	New Zealand
Canada	Nigeria
East Africa (Kenya, Malawi, Tanzania)	Pakistan
Fiji	Philippines
Great Britain (parsed)	Sierra Leone
Hong Kong	Singapore
India	South Africa
Ireland	Sri Lanka
Jamaica	Trinidad and Tobago
Kenya	USA
Malta	

ICE-GB was first released in 1998 with ICECUP 3.0. Since then it has been used for research and education in universities, colleges and schools all over the world.



Video corpora

<http://sourceforge.net/projects/thedrs>

Home / Browse / Audio & Video / Video / Display / Digital Replay System



Digital Replay System

Status: **Beta** Brought to you by: afrench, cgreenhalgh, chaoticgallen, janhumble, pxt

Add a Review Downloads: 1 This Week Last Update: 2013-04-15

 Download Get Updates Share This

Mac | Windows

Summary Files Reviews Support Wiki Tickets ▾ News Discussion Code

The DRS (Digital Replay System) is a desktop application for replaying and analysing combinations of video, audio, images, transcripts and computer log files in an integrated way.

1.2. Corpus-based versus corpus-driven linguistics

Corpus-based

- corpus linguistics is perceived as a **methodology** ⇒ corpus data are used *to verify the existing theories of language*

Corpus-driven

- tends to view corpus linguistics as a **theory** which offers a new way of looking at the creation of meaning in a narrow sense and different aspects of the use of language in a broader sense

Corpus-based studies

use corpus data in order to
explore a theory or hypothesis, established in
the current literature, in
order to validate it, refute
it or refine it

[McEnery & Hardie 2012: 6]

Corpus-driven linguistics

- claims that the corpus itself should be the sole source of hypotheses about language
- the corpus itself *embodies its own theory of language*

[McEnery & Hardie 2012: 6].

1.3. Data collection regime

The monitor corpus approach

- seeks to develop a **dataset** which ***grows in size over time*** and which contains a ***variety of materials***

The Bank of English (BoE)

many books on corpus linguistics suggested that the BoE could be used as a '*monitor corpus*' to look ongoing changes in English



The Web as Corpus

- It takes as its starting point a massive collection of data that is *ever-growing*, and uses it for the study of language.
- The content of the web is not ***divided by genre*** ⇒ the material returned from a web search tends to be an ***undifferentiated mass***, which requires a ***great deal of processing*** to sort into meaningful groups of texts.



The sample corpus approach

- ✓ The *sample corpora* represent a particular type of language over a specific span of time
- ✓ A *balanced corpus* covers a wide range of text categories which are supposed to be *representative of the language variety* under consideration.
- ✓ *Representativeness* refers to the extent to which a sample includes the *full range of variability* in a population.



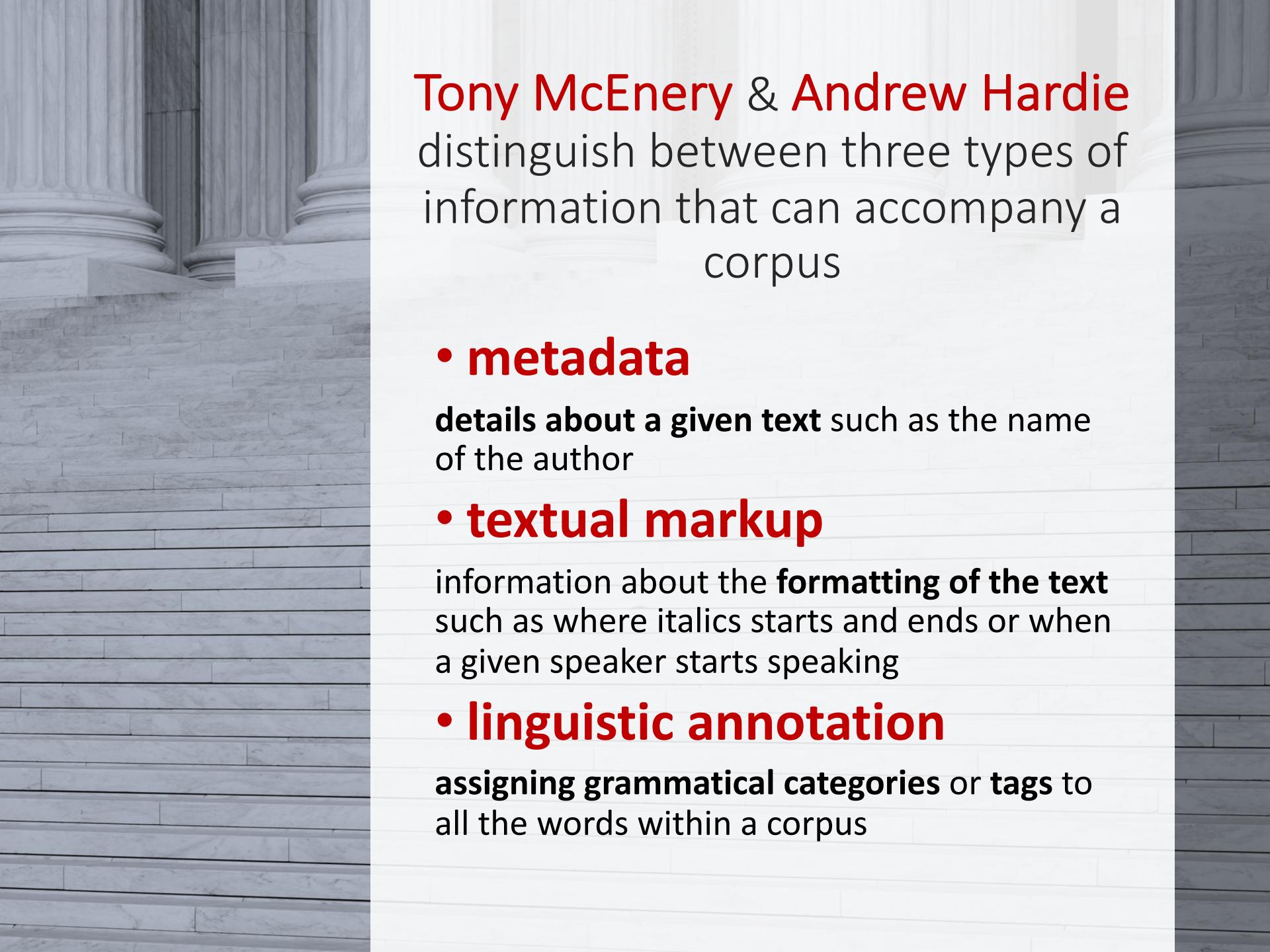
1.4. Annotated versus unannotated corpora



Corpus annotation is largely the process of providing those analyses which a linguist would carry out anyway on whatever data they worked with.



Annotation is an umbrella term that refers to procedures such as **tagging** and **parsing** which are carried out to add linguistic information to a corpus



Tony McEnergy & Andrew Hardie
distinguish between three types of
information that can accompany a
corpus

- **metadata**

details about a given text such as the name
of the author

- **textual markup**

information about the **formatting of the text**
such as where italics starts and ends or when
a given speaker starts speaking

- **linguistic annotation**

assigning grammatical categories or tags to
all the words within a corpus



Layers of annotation

- *part-of-speech (PoS) tagging*
- *syntactic (grammatical) parsing*
- *error annotation*
- *semantic annotation*
- *phonetic annotation*

CLAWS

Constituent Likelihood Automatic Word-tagging System

<http://ucrel.lancs.ac.uk/claws/>

CLAWS part-of-speech tagger for English

[Free CLAWS WWW tagger](#) | [Obtaining a licence](#) | [Tagging service](#)

Introduction

Part-of-speech (POS) tagging, also called grammatical tagging, is the commonest form of [corpus annotation](#), and was the first form of annotation to be developed by [UCREL](#) at Lancaster. Our POS tagging software for English text, CLAWS (the Constituent Likelihood Automatic Word-tagging System), has been continuously developed since the early 1980s. The latest version of the tagger, CLAWS4, was used to POS tag c.100 million words of the [British National Corpus](#) (BNC).

Accuracy

CLAWS has consistently achieved 96-97% accuracy (the precise degree of accuracy varying according to the type of text). Judged in terms of major categories, the system has an error-rate of only 1.5%, with c.3.3% ambiguities unresolved, within the BNC. More detailed analysis of the error rates for the C5 tagset in the BNC can be found within the [BNC manual](#).

Template tagging

In the context of the BNC Enhancement project, UCREL devised a Template Tagger to act as a post processor for CLAWS. The rule-based formalism implemented in the Template Tagger is more powerful than that built into CLAWS itself. Manual corpus analysis and knowledge of frequent CLAWS tagging errors was used to create a rule base for the tool. This facilitated an improvement in the tagging accuracy in the resulting corpus. For more details, see [Fligelstone, Rayson, and Smith \(1996\)](#) and [Fligelstone, Pacey, and Rayson \(1997\)](#). Please note that the Template Tagger processing is not currently included in the online tagger or the licenced versions of CLAWS4. Please contact us for further details of current availability.

Tagging services

UCREL offers access to our latest version of CLAWS4 by:

- [selling site and single user licences](#) for software use within academic institutions and commercial organisations
- providing an in-house [tagging service](#) at Lancaster University
- our free [CLAWS WWW tagger](#) where you can submit text to be POS tagged via the Internet
- CLAWS can also be accessed through the web-based [Wmatrix interface](#)

1.5. Total accountability versus data selection

The principle of
**total
accountability**

we must not
select a
favourable subset
of the data

one way of
satisfying
falsifiability is to
**use the entire
corpus** to test the
hypothesis

1.6. Multilingual versus monolingual corpora

Many corpora are **monolingual**

- may represent a range of varieties and genres of a particular language
- limited to that one language

Survey of English Usage

- ➊ Home
- ➋ About the Survey
- ➌ Staff
- ➍ Research Projects
 - ➎ The International Corpus of English (ICE)
 - ➏ ICE-GB
 - ➐ DCPSE
 - ➑ Corpus Queries
 - ➒ Next Generation Tools
 - ➓ The English Noun Phrase
 - ➔ The English Verb Phrase
 - ➕ Subordination in Spoken & Written English
 - ➖ Teaching English Grammar in Schools
- ➗ Research Resources
- ➘ Software Sales
- ➙ Mobile Apps
- ➚ Events
- ➛ Summer School
- ➜ Study With Us
- ➝ Courses for Teachers
- ➞ Survey Archives

The International Corpus of English

The **International Corpus of English** (ICE) project was initiated in 1988 by the late Sidney Greenbaum, the then Director of the Survey of English Usage, University College London. In a brief notice in *World Englishes*, Greenbaum pointed out that grammatical studies had been greatly facilitated by the availability of two computerized corpora of printed English, the Brown Corpus of American English, and the LOB (Lancaster/Oslo-Bergen) Corpus of British English. Greenbaum continued:

We should now be thinking of extending the scope for computerized comparative studies in three ways: (1) to sample standard varieties from other countries where English is the first language, for example Canada and Australia; (2) to sample national varieties from countries where English is an official additional language, for example India and Nigeria; and (3) to include spoken and manuscript English as well as printed English. (Greenbaum 1988)

In response, linguists from around the world came forward to discuss Greenbaum's proposal, and ultimately to put it into effect (Greenbaum 1991). The project soon became known as the International Corpus of English (ICE), and was coordinated by Greenbaum until 1996. From 1996 to 2001, ICE was coordinated by Charles Meyer, University of Massachusetts-Boston. It is now coordinated by Gerald Nelson in Hong Kong. The ICE project involves research teams in each of the countries or regions shown below.

Australia	Malaysia
Cameroon	New Zealand
Canada	Nigeria
East Africa (Kenya, Malawi, Tanzania)	Pakistan
Fiji	Philippines
Great Britain	Sierra Leone
Hong Kong	Singapore
India	South Africa
Ireland	Sri Lanka
Jamaica	Trinidad and Tobago
Kenya	USA



➊ ICE website



➏ ICE-GB

➐ DCPSE

The English-Norwegian Parallel Corpus

Sub-corpora

Team

The English-
Norwegian Parallel
Corpus

The English-Norwegian Parallel Corpus (ENPC) consists of original texts and their translations (English to Norwegian and Norwegian to English).

It is intended as a general research tool, available beyond the present project for applied and theoretical linguistic research. It started out as a research project at the Department of British and American Studies, University of Oslo, in 1994. The corpus was completed in 1997. In the period 1997-2001 the corpus was extended to include more languages (German, Dutch, Portuguese), and the English and the Norwegian original texts were tagged for part of speech. The manual was completed in 1999 and revised in 2002.

The focus has been on novels and fairly general non-fictional books. In order to include material by a range of authors and translators, the texts of the corpus are limited to text extracts (chunks of 10,000-15,000 words). The fiction part of the corpus contains 30 original text extracts in each language and their translations, whereas the non-fiction part contains 20 in each direction.

the English-Norwegian Parallel Corpus (ENPC)

<https://www.hf.uio.no/ilos/english/services/knowledge-resources/omc/enpc/>

Type A: Source texts in one language plus translations into one or more other languages

- *the Canadian Hansard*

- consisting of **debates from the Canadian Parliament** published in the country's official languages, *English* and *French*

- *CRATER*

- **Corpus Resources and Terminology Extraction** is a project involving three languages: *English*, *French* and *Spanish*.
- consists entirely of **technical texts** from the International Telecommunications Union ⇒ 5,5 million words
- texts are tagged with part-of-speech and morphological annotation

Type B: Pairs or groups of monolingual corpora designed using the same sampling frame

<https://www.lancaster.ac.uk/fass/projects/corpus/LCMC/>

The Lancaster Corpus of Mandarin Chinese

(LCMC)

by

Tony McEnery

Richard Xiao

Lancaster University

Preface

The Lancaster Corpus of Mandarin Chinese (LCMC) addresses an increasing need within the research community for a publicly available balanced corpus of Mandarin Chinese. LCMC has been constructed as part of a research project undertaken by the Linguistics Department, Lancaster University. The corpus is designed as a Chinese match of the *Freiburg-LOB Corpus of British English* (*FLOB*), and, as such, will provide a valuable resource for contrastive studies between English and Chinese as well as a sound basis for monolingual investigations of Chinese. The LCMC corpus is distributed by the [European Language Resources Association](#) (Cat. No ELRA-W0039) and the [Oxford Text Archive](#) (Cat. No 2474).

We are obliged to the **UK Economic and Social Research Council** for funding our project (see Grant Ref. RES-000-220135). Without their help, this corpus would not have been built. We would also like to thank the presses, libraries and websites, as listed in the bibliographic document of this corpus, for providing the required texts, and Miss Xin Huang, for proofreading the scanned electronic texts.

Type C: A combination of A and B

EMILLE

Enabling Minority Language Engineering

was a 3-year project at Lancaster University and
Sheffield University

Its end product was a 97 million word electronic
corpus of **South Asian languages**, especially those
spoken in the UK

<http://www.emille.lancs.ac.uk/about.php>

2. Providing data on linguistic phenomena

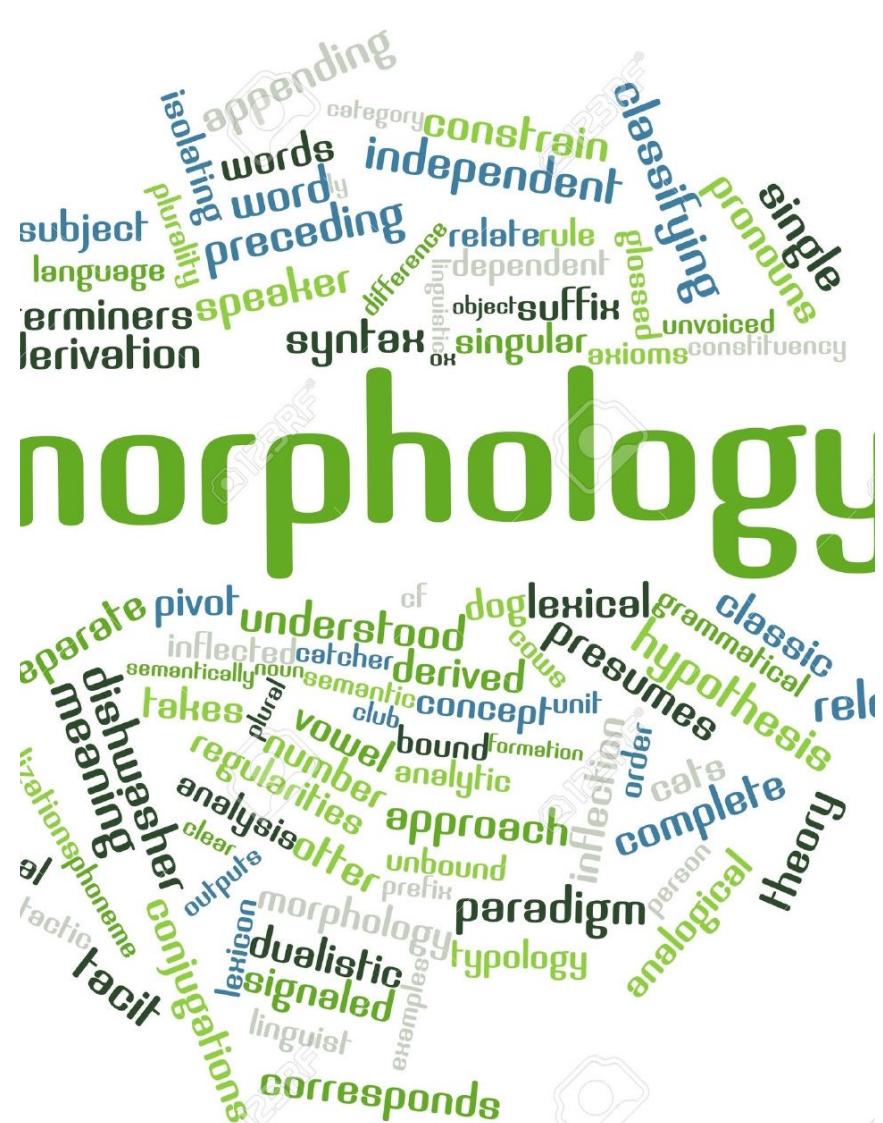
Lexical

Frequency and distribution of specific words and phrases

Lists of all common words in a language or genre

- processes involving **word formation**
nouns formed with
suffixes ***ism** or
***ousness**
 - contrasts in the use
of grammatical
alternatives

HAVE + *proven/proved*
sincerest/most sincere





High-frequency grammatical features ↳ *modals, passives, perfect or progressive aspect*

↳ **Less frequent** grammatical variation

John started to walk / walking

She'd like (for) him to stay overnight

Phraseological patterns

- **Collocational preferences**

for specific words

true feelings, true story

- **Constructions**

[V NP into V-ing]

they talked him into staying

[V POSS way PREP]

*he elbowed his way through
the crowd*

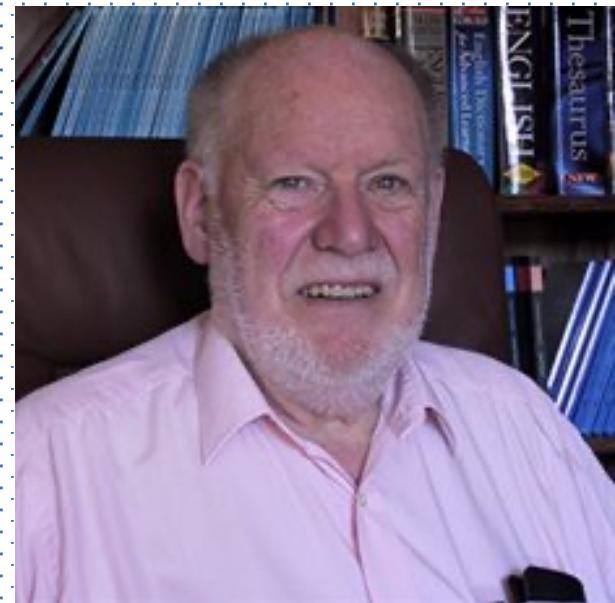


SEMANTICS

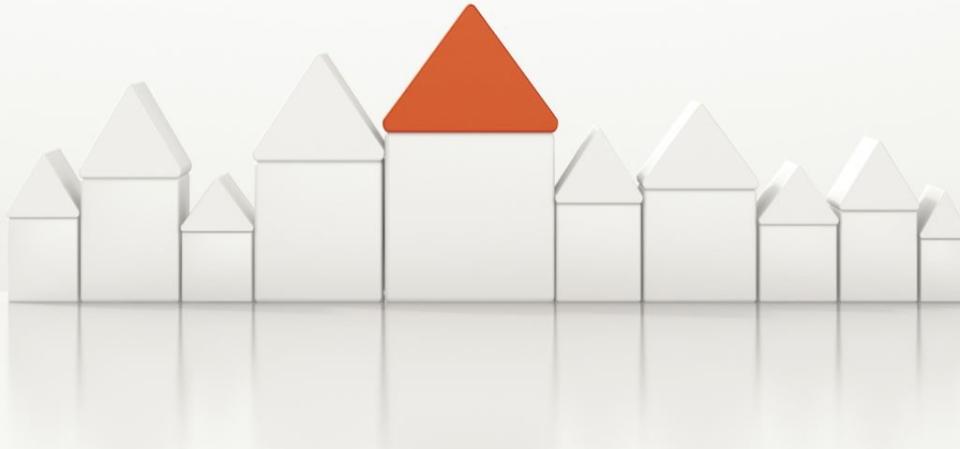
- **Collocates** as a guide to meaning and usage
(n) *highbrow*, (adj) *highbrow*
- **Semantic prosody**
the types of words preceding the verb
outweigh

3. Corpus design

In 2005, Sinclair proposed a set of principles that should be considered with regard to the process of **developing a corpus**



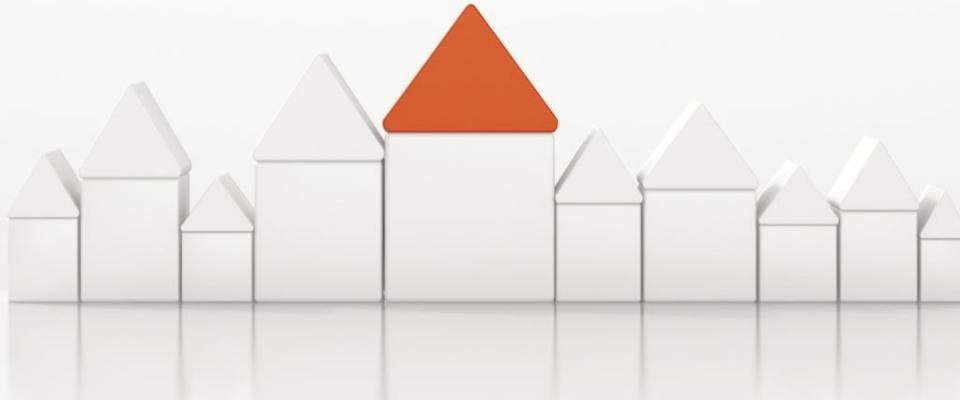
John Sinclair
(1933-2007)



- 1. The **contents of a corpus** should be selected without regard for the language they contain, but according to their **communicative function** in the community in which they arise.
- 2. Corpus builders should strive to make their corpus as **representative** as possible of the language from which it is chosen.
- 3. Only those components of corpora which have been designed to be **independently contrastive** should be contrasted.

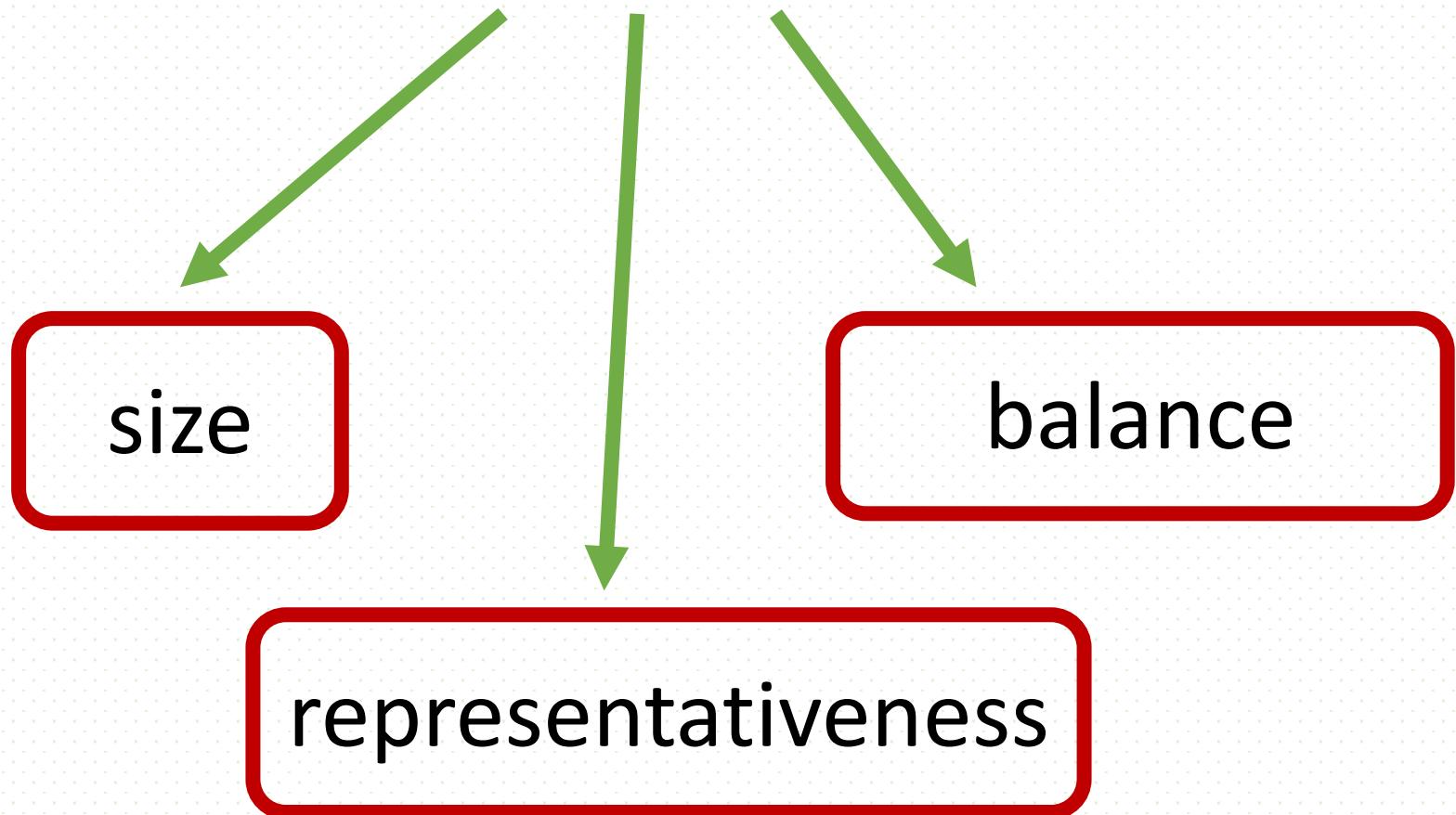


- 4. **Criteria** for determining the structure of a corpus should be *small in number*, clearly *separate* from each other and *efficient as a group* in delineating a corpus that is representative of the language variety under examination.
- 5. Any *information about a text* other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.
- 6. Samples of language for a corpus should consist of *entire documents* or transcriptions of complete speech events.



- 7. The **design and composition** of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.
- 8. The corpus builder should retain, as target notions, **representativeness** and **balance**. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.
- 9. Any control of subject matter in a corpus should be imposed by the use of **external**, and not internal, **criteria**.
- 10. A corpus should aim for **homogeneity** in its components while maintaining adequate coverage, and rogue texts should be avoided.

Corpus research



- **Representativeness** concerns the issue of **how well a corpus represents a given language** or variety that is under study.
- **Balance** refers to the **structure and type of data** used to build a corpus.
- A *well-balanced corpus* should consist of several subsections that represent different types of language use.

The COCA corpus (new version released March 2020)

The corpora from English-Corpora.org are the world's **most widely-used corpora**. The Corpus of Contemporary American English (COCA) is by far the most widely-used of these corpora. In early 2020, we dramatically expanded the scope and size and features of COCA to make it even more useful for researchers, teachers, and learners.

The corpus contains more than **one billion words** of data, including 20 million words each year from **1990-2019** (with the same genre balance year by year). This makes COCA the only corpus of English that is 1) large 2) recent and 3) has a wide range of genres. The following table shows the genres in the corpus.

Genre	# texts	# words	Explanation
Spoken	44,803	127,396,932	Transcripts of unscripted conversation from more than 150 different TV and radio programs (examples: All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC), Oprah)
Fiction	25,992	119,505,305	Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990-present, and fan fiction.
Magazines	86,292	127,352,030	Nearly 100 different magazines, with a good mix between specific domains like news, health, home and gardening, women, financial, religion, sports, etc.
Newspapers	90,243	122,958,016	Newspapers from across the US, including: USA Today, New York Times, Atlanta Journal Constitution, San Francisco Chronicle, etc. Good mix between different sections of the newspaper, such as local news, opinion, sports, financial, etc.
Academic	26,137	120,988,361	More than 200 different peer-reviewed journals. These cover the full range of academic disciplines, with a good balance among education, social sciences, history, humanities, law, medicine, philosophy/religion, science/technology, and business
Web (Genl)	88,989	129,899,427	Classified into the web genres of academic, argument, fiction, info, instruction, legal, news, personal, promotion, review web pages (by Serge Sharoff). Taken from the US portion of the GloWbE corpus.
Web (Blog)	98,748	125,496,216	Texts that were classified by Google as being blogs. Further classified into the web genres of academic, argument, fiction, info, instruction, legal, news, personal, promotion, review web pages. Taken from the US portion of the GloWbE corpus.
TV/Movies	23,975	129,293,467	Subtitles from OpenSubtitles.org, and later the TV and Movies corpora. Studies have shown that the language from these shows and movies is even more colloquial / core than the data in actual "real speech".

4. Benefits of corpus analysis

- 1. One can use corpus data to explore *different aspects of language*.
- 2. Corpus linguistics is an empirical approach which relies on *frequency-based analyses*.
- 3. Corpus linguistics focuses on the *phraseological nature of language*.
- 4. Corpus investigations highlight different *functions* of language and demonstrate the *central role of context* in the analysis of linguistic behavior.
- 5. Corpus linguistics presents us with powerful tools for exploring the *distribution of specific linguistic features* across a wide range of domains of language use.

5. Limitations of corpus analysis

1. A corpus can show us only *what it contains*.
2. A corpus may be *too small*.
3. A corpus presents language *out of its context*.
4. A corpus *cannot interpret data*.

A yellow bomb icon is positioned in the top-left corner of the slide. It consists of a large yellow circle with a smaller yellow circle inside it, and three yellow dashes radiating from the top left of the inner circle.

6. Types of corpora

6.1. General and specialized corpora

General corpora consist of a wide range of texts that represent natural language as it is used *across a variety of contexts*.

Specialized corpora do not aim to comprehensively represent a language as a whole, but only *specialized segments* of it.



SEARCH

FREQUENCY

CONTEXT

OVERVIEW

[List](#) [Chart](#) [Collocates](#) [Compare](#) [KWIC](#) [POS]?[Find matching strings](#)[Reset](#) Sections [Texts/Virtual](#) [Sort/Limit](#) [Options](#)

(HIDE HELP)

+ LICENSE

The [British National Corpus \(BNC\)](#) was originally created by [Oxford University press](#) in the 1980s - early 1990s, and it contains [100 million words](#) of text from a wide range of genres (e.g. spoken, fiction, magazines, newspapers, and academic).

The BNC is related to many other [corpora of English](#) that we have created. These corpora were formerly known as the "BYU Corpora"), and they offer unparalleled insight into [variation in English](#).

Click on any of the links in the search form to the left for context-sensitive help, and to see the range of queries that the corpus offers. You might pay special attention to the [comparisons between genres](#) and [virtual corpora](#), which allow you to create personalized collections of texts related to a particular area of interest.

BNC <https://www.english-corpora.org/bnc/>



Michigan Corpus of Academic Spoken English

Welcome to our NEW interface to the on-line, searchable part of our collection of transcripts of academic speech events recorded at the University of Michigan.

There are currently 152 transcripts (totaling 1,848,364 words) available at this site.

Browse MICASE

Browse the corpus according to specified speaker and speech attributes, returning quick file references.

Search MICASE

Search the corpus for words or phrases in specified contexts, returning concordance results with references to files, full utterances, and speakers.

Michigan Corpus of Academic Spoken English
(MICASE) <https://quod.lib.umich.edu/m/micase/>

British Academic Written English Corpus



 Please use the following text to cite this item or export to a predefined format:

[BIBTEX](#)[CMDI](#)

Nesi, Hilary; Gardner, Sheena; Thompson, Paul; et al., 2008, *British Academic Written English Corpus*, Oxford Text Archive,
<http://hdl.handle.net/20.500.12024/2539>.



 Share: [f](#) [t](#) [g](#)

[Oxford Text Archive](#)

 Authors	Nesi, Hilary ; Gardner, Sheena ; Thompson, Paul ; Wickens, Paul
 Date of publication	2004
 Type	Corpus
 Language(s)	English
 OTA identifier	ota:2539
 Collection(s)	Core Collection

OXFORD
TEXT
ARCHIVE



Bodleian Libraries
UNIVERSITY OF OXFORD

 Browse

> All of the Repository

 My Account

 Login

 Statistics

 Statistics

BETA

 General Information

 Cite

 Oxford University users

 FAQ

British Academic Written English (BAWE) Corpus of proficient student writing

<https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2539#>



Vienna-Oxford International Corpus of English

About VOICE

[What is VOICE](#)
[Corpus Description](#)
[FAQ](#)
[News](#)
[Thanks](#)

Corpus

[Availability](#)
[Access VOICE Online](#)
[Corpus Information](#)
[Corpus Compilers](#)
[Using VOICE Online](#)
[Corpus Statistics](#)
[Download VOICE XML](#)
[Citing VOICE](#)
[Terms of Use \(VOICE Online\)](#)

Research

[Research Perspectives](#)
[Publications & Other Studies](#)
[Presentations](#)
[VOICE-based Publications](#)

Transcription and Annotation

[General information on the VOICE transcription conventions](#)
[Mark-up Conventions](#)
[Spelling Conventions](#)
[VoiceScribe](#)
[Tagging and Lemmatization](#)

[← home](#)

What is VOICE?

A computer corpus of English as a lingua franca

The most wide-spread contemporary use of English throughout the world is that of **English as a lingua franca (ELF)**, i.e. English used as a common means of communication among speakers from different first-language backgrounds. A Hungarian educationalist coming to Copenhagen to discuss qualification equivalences in European higher education with her Danish, Finnish and Portuguese colleagues; a Korean sales representative negotiating a contract with his German client in Luxembourg; a Spanish Erasmus student chatting with local colleagues in a student hall in Vienna: they all communicate in English as a lingua franca.

VOICE, the Vienna-Oxford International Corpus of English, is a structured collection of language data, the first computer-readable corpus capturing spoken ELF interactions of this kind.

VOICE, compiled at the Department of English at the University of **Vienna**, is funded by the **Austrian Science Fund** (FWF). These funds were further supplemented by a contribution from Oxford University Press in 2008. Supporting funds were also provided in the early pilot phase by Oxford University Press and by the Hochschuljubiläumsstiftung der Stadt Wien. The corpus currently comprises 1 million words of transcribed spoken ELF from professional, educational and leisure domains.

*Vienna-Oxford International Corpus of English,
VOICE (<https://www.univie.ac.at/voice/>)*

English as a Lingua Franca in Academic Settings Corpus, ELFA. <https://www.kielipankki.fi/corpora/elfa/>



The screenshot shows the main navigation bar with links: LANGUAGE BANK, ACCESS, CORPORA, TOOLS, ORGANIZATION, SUPPORT, SUOMEKSI, and PÅ SVENSKA.

ELFA – English as a Lingua Franca in Academic Settings

Current versions of this resource:

- The Helsinki Korp Version of the ELFA Corpus**
Select the corpus in Korp ?
Metadata and license
Attribution instructions
- The Transcriptions of the ELFA Corpus, Downloadable Version**
Download the resource

Search the Language Bank Portal:

 Researcher of the Month: Juho Leinonen

6.2. Written and spoken corpora

The majority of corpora represent written language

CANCODE (*Cambridge and Nottingham Corpus of Discourse in English*) Corpus

large collection of spoken British English and it has been used as a basis for a number of studies into the specific nature of spoken language

Santa Barbara Corpus of Spoken American English

<https://www.linguistics.ucsb.edu/research/santa-barbara-corpus>



Home People Research Graduate Undergraduate Courses Resources News & Events Alumni Giving

RESEARCH MENU

- [Overview](#)
- [Corpus Linguistics](#)
- [Discourse & Grammar](#)
- [Language & Cognition](#)
- [Language Change](#)
- [Language Documentation](#)
- [Prosody](#)
- [Sociocultural Linguistics](#)
- [Typology](#)

- [Applied Linguistics](#)
- [Language Areas](#)
- [Transcription](#)
- [Pear Film World Corpus](#)
- [Santa Barbara Corpus of Spoken American English](#)
- [Santa Barbara Papers in Linguistics](#)

Santa Barbara Corpus of Spoken American English

Parts 1-4 of the Santa Barbara Corpus of Spoken American English (SBCSAE) are now available, for a total of approximately 249,000 words. The Santa Barbara Corpus includes transcriptions, audio, and timestamps which correlate transcription and audio at the level of individual intonation units.

- [Access](#)
- [Description](#)
- [Contents and Summaries](#)
- [Citation](#)
- [Recordings](#)
- [Acknowledgements](#)
- [Contact](#)

Access

All transcriptions in the Santa Barbara Corpus parts 1-4 can be downloaded for free by clicking [here](#). Metadata is available [here](#).

To access individual conversations and other discourse segments in the Santa Barbara Corpus, you may select the audio file and transcription you wish to download by consulting the [Contents and Summaries](#).

To download the audio files in WAV (recommended) or MP3 format, do the following:

Hong Kong Corpus of Spoken English

<http://rcpce.engl.polyu.edu.hk/HKCSE/default.htm>

The screenshot shows the homepage of the Hong Kong Corpus of Spoken English (HKCSE). At the top left is the logo for the Research Centre for Professional Communication in English (RCPCE), featuring the letters 'RC' and 'PCE' in blue and green. To its right is the logo for The Hong Kong Polytechnic University, which includes a red square emblem and the text 'THE HONG KONG POLYTECHNIC UNIVERSITY' and '香港理工大學'. Further right is the logo for the Department of English and Communication, featuring a blue stylized 'i' icon and the text 'Department of English and Communication' and '英文及傳意學系'. Below these logos is a navigation bar with the text 'RCPCE Profession-specific Corpora' and a dropdown menu containing the instruction 'Click here to select a profession-specific corpus'. To the right of the dropdown is a link 'Back to Front Page' and the LinkedIn logo. The main title 'Hong Kong Corpus of Spoken English' is centered below the navigation bar. A paragraph of text welcomes visitors to the HKCSE, stating it is a large collection of texts representing spoken English in Hong Kong, available in orthographic and prosodic versions. It also links to the John Benjamins website. Below this text is a section for citation information, listing a publication by Cheng W., Greaves C., and Warren M. (2005) in the ICAME Journal. A note indicates there are currently 907,657 words in the corpus. Finally, a list of features is provided, including search options for words, phrases, and additional words/phrases.

RC Research Centre for
PCE Professional Communication in English

THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

Department of English
and Communication
英文及傳意學系

RCPCE Profession-specific Corpora

Click here to select a profession-specific corpus

Back to Front Page

in®

Hong Kong Corpus of Spoken English

Welcome to the [Hong Kong Corpus of Spoken English \(HKCSE\)](#) hosted by the Research Centre for Professional Communication in English of the Hong Kong Polytechnic University. The HKCSE is a large collection of texts representing spoken English in Hong Kong. This is the orthographic version, if you would like to purchase or know more about the prosodic version (A corpus-driven study of discourse intonation with a CD). Click here to go to [John Benjamins website](#).

Please cite the HKCSE with following information:

Cheng W, Greaves C, Warren M (2005). The creation of prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic), *ICAME Journal*, vol. 29 (pg. 47-68), April 2005.

There are currently **907,657** words in the HKCSE.

- You can search for a word, e.g. people, not, or a phrase, e.g. Hong Kong people, a lot of, and find examples of its use in its context.
- You can also search for an additional word in combination with your search word, e.g. people (search word) and different (additional word), or search phrase, e.g. I don't know (search phrase) and actually(additional word).

6.3. Historical (diachronic) corpora

- Represent data from ***specific historical periods*** and they are particularly useful if scholars are interested in the process of ***language change***

Corpus of Historical American English, COHA

<https://www.english-corpora.org/coha/>

The screenshot shows the COHA search interface. On the left, there's a search form with fields for 'Text' and 'POS?', and buttons for 'Find matching strings' and 'Reset'. Below this are links for 'List', 'Chart', 'Collocates', 'Compare', and 'KWIC'. A checkbox for 'Sections' is checked. On the right, there's a 'SEARCH' tab, a 'FREQUENCY' tab, a 'CONTEXT' tab, and an 'ACCOUNT' tab. Above the tabs are icons for help, download, and user profile. To the right of the search form is a large text area containing information about the COHA corpus, its size, and its relation to other corpora like Google Books and Hansard. It also includes a link to download the corpus for offline use and a note about updates in 2021.

Corpus of Historical American English

SEARCH FREQUENCY CONTEXT ACCOUNT

List Chart Collocates Compare KWIC

[Text] [POS]? Find matching strings Reset

Sections Texts/Virtual Sort/Limit Options

(HIDE HELP) + LICENSE

Download the corpus for offline use

Updates and enhancements in 2021

The Corpus of Historical American English (COHA) is the largest structured corpus of historical English. It is related to many other [corpora of English](#) that we have created. These corpora were formerly known as the "BYU Corpora", and they offer unparalleled insight into [variation in English](#). If you are interested in historical corpora, you might also look at our [Google Books](#) (see [comparison](#)), [Hansard](#), and [TIME](#) corpora.

COHA contains more than [475 million words](#) of text from the 1820s-2010s (which makes it [50-100 times as large](#) as other comparable historical corpora of English) and the corpus is balanced by genre decade by decade. The creation of the corpus results from a grant from the [National Endowment for the Humanities](#) (NEH) from 2008-2010.

Click on any of the links in the search form to the left for context-sensitive help, or take a look at the [expanded overview](#) of the corpus. You might pay special attention to the [comparisons between decades](#) and [virtual corpora](#), which allow you to create personalized collections of texts related to a particular area of interest.

ARCHER, A Representative Corpus of Historical English Registers

<https://www.projects.alc.manchester.ac.uk/archer/>



ARCHER: A Representative Corpus of Historical English Registers

[ARCHER versions](#)[Consortium members](#)[Using ARCHER](#)[Publications](#)[Documentation](#)

ARCHER: A Representative Corpus of Historical English Registers

ARCHER is a multi-genre corpus of British and American English covering the period 1600-1999, first constructed by Douglas Biber and Edward Finegan in the 1990s. It is managed as an ongoing project by a [consortium of participants](#) at fourteen universities in seven countries. On receipt of an electronically signed User Agreement (print version downloadable from the [Documentation](#) page of this site), ARCHER will be made available both for in-house use at the consortium universities and for online internet searches. An instructor (lecturer, teacher) may submit a User Agreement on behalf of a class of students.

For hints and tips for registered users, please visit the last section of the [Using ARCHER](#) page.

For a list of publications that have made use of ARCHER at one of the participating departments, please visit [Publications](#) page.

 Search

6.4. Parallel and comparable corpora

- Often employed by researchers working in the area of ***translation studies***, who use them to make direct comparisons between the same texts written in different languages

Oslo Multilingual Corpus

<http://www.hf.uio.no/ilos/english/services/omc/>

German, French and Finnish source texts, and their respective translations



Faculty of Humanities

Department of Literature, Area Studies and European Languages

☰ Menu

← Services and tools

← Knowledge resources

[Norwegian
version of this
page](#)

Oslo Multilingual
Corpus

▪ Sub-corpora

▪ Team

▪ The English-
Norwegian
Corpus

Oslo Multilingual Corpus - background and use

The Oslo Multilingual Corpus (OMC) is a collection of text corpora comprising original texts and translations from several languages.

The various sub-corpora differ in that they contain a different number of languages or a different combination of languages.

En akademisk

Digital Corpus of the European Parliament

<https://ec.europa.eu/jrc/en/language-technologies/dcep>

The screenshot shows the EU Science Hub website. At the top, there is a banner with the European Commission logo and the text "An official website of the European Union" and "How do you know?". The main navigation bar includes links for Home, About Us, Research, Knowledge, Working With Us, Procurement, News & Events, and Our Communities. Below the navigation, a sidebar lists various language technology resources: JRC-Acquis, DGT-Acquis, DCEP: Digital Corpus of the European Parliament (highlighted in blue), DGT Translation Memory, ECDC Translation Memory, EAC Translation Memory, JRC-Names, and JRC Eurovoc Indexer. The main content area features a large heading "DCEP: Digital Corpus of the European Parliament" and a bulleted list of topics including Introduction, Format and Structure of the Data, Document types contained in DCEP, Statistics on the DCEP, Usage conditions, Download the DCEP corpus, How to produce bilingual corpora, Acknowledgement and contact, Reference publication, and International Standard Language Resource Number: 823-807-024-162-2. To the right, a "Related Content" sidebar lists links to Competence Centre on Text Mining and Analysis, Language Technology Resources, Scientific Publications by the EMM Team, Europe Media Monitor - NewsBrief, Europe Media Monitor - NewsExplorer, Medical Information System (MedISys), Tools for Innovation Monitoring, and EMM App for mobile devices.

An official website of the European Union How do you know?

English (en) ▾

Search

EU SCIENCE HUB
The European Commission's science and knowledge service

European Commission > EU Science Hub > Language Technology Resources > Dcep

Home About Us Research Knowledge Working With Us Procurement News & Events Our Communities

Language Technology Resources

- JRC-Acquis: a multilingual parallel corpus
- DGT-Acquis: multilingual parallel corpora from the EU Official Journal
- DCEP: Digital Corpus of the European Parliament**
- DGT Translation Memory
- ECDC Translation Memory
- EAC Translation Memory
- JRC-Names: names and their spelling variants
- JRC Eurovoc Indexer

DCEP: Digital Corpus of the European Parliament

- Introduction
- Format and Structure of the Data
- Document types contained in DCEP
- Statistics on the DCEP
- Usage conditions
- Download the DCEP corpus
- How to produce bilingual corpora
- Acknowledgement and contact
- Reference publication
- International Standard Language Resource Number: 823-807-024-162-2

Related Content

- [Competence Centre on Text Mining and Analysis](#)
- [Language Technology Resources](#)
- [Scientific Publications by the EMM Team](#)
- [Europe Media Monitor - NewsBrief](#)
- [Europe Media Monitor - NewsExplorer](#)
- [Medical Information System \(MedISys\)](#)
- [Tools for Innovation Monitoring](#)
- [EMM App for mobile devices](#)

6.5. Web as a corpus

Internet is **constantly growing** and the number of websites is increasing.

A corpus can be regularly **updated**, which makes it very similar to monitor corpora.

WebCorp Linguist's Search Engine

<http://wse1.webcorp.org.uk/>

is an example of an interface that can be used to explore data found on the web

The WebCorp Linguist's Search Engine is a tool for the study of language on the web. The corpora below were built by crawling the web and extracting textual content from web pages. Searches can be performed to find words or phrases, including pattern matching, wildcards and part-of-speech. Results are given as concordance lines in KWIC format. Post-search analyses are possible including time series, collocation tables, sorting and summaries of meta-data from the matched web pages.

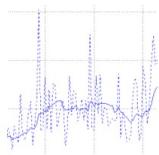
A snippet of text from a web page, showing occurrences of the word 'research'.

Synchronic English Web Corpus

470 million word corpus built from web-extracted texts. Including a randomly selected 'mini-web' and high-level subject classification.

[Search ▶](#)

[About](#)



Diachronic English Web Corpus

130 million word corpus randomly selected from a larger collection and balanced to contain the same number of words per month.

[Search ▶](#)

[About](#)

lovely	2057
me	1846
another	1633
beautiful	1407
wonderful	1444
started	1218
discovered	1198
enjoy	1052
see	504

A snippet of text from a personal letter, showing French words and some English words.

Birmingham Blog Corpus

630 million word corpus built from blogging websites. Including a 180 million word sub-section separated into posts and comments.

[Search ▶](#)

[About](#)



Anglo-Norman Correspondence Corpus

A corpus of approximately 150 personal letters written by users of Anglo-Norman. Including bespoke part-of-speech annotation.

[Search ▶](#)

[About](#)

Novels of Charles Dickens

A searchable collection of the novels of Charles Dickens. Results can be visualised across chapters and novels.

[Search ▶](#)

[About](#)

Mark Davies's Google Books interface

<https://support.google.com/websearch/answer/9523832>

<https://books.google.com/ngrams/info>

Use the new Google Books

You can read, download, or preview books on Google Books. If you find a book you want to read, you might be able to read it on Google Books, buy it online, or borrow it from a library.

Tip: Some books are provided by publishers, while others are scanned as part of the [Library Project](#).

Important: Some of these features might not be available on mobile devices.

Try the new Google Books

You can try the [new Google Books here](#) or:

1. Go to [Google Books](#).
2. Search for any book.
3. Click the book title.
4. At the top left, click Try it now.

Go back to classic Google Books

You can [go back to classic Google Books here](#) from your computer or:

1. Go to [Google Books](#).
2. Search for any book.
3. Click the book title.
4. If you are viewing pages of a book, click Close .
5. At the top, click Settings .
6. Click Go back to classic Google Books.

Use Google Books

-  [How to use classic Google Books](#)
-  [Add, organize, or share books](#)
-  [Use the new Google Books](#)
-  [About the Library Project](#)
-  [Claim or exclude your books in the Library Project](#)

The example of a database composed of web-based data is **the NOW Corpus**

<https://www.english-corpora.org/now/>

The screenshot shows the homepage of the NOW Corpus. At the top, there's a blue header bar with the title "NOW Corpus (News on the Web)" and icons for help, file, download, user profile, and settings. Below the header are four tabs: "SEARCH" (selected), "FREQUENCY", "CONTEXT", and "OVERVIEW". The main content area has two columns. The left column contains a search form with fields for "List", "Chart", "Collocates", "Compare", and "KWIC". It also includes a text input field, a dropdown for "[POS]?", and buttons for "Find matching strings" and "Reset". Below this is a section for "Search by date" with a checkbox for "Sections" and links for "Texts/Virtual", "Sort/Limit", and "Options". The right column starts with a "HIDE HELP" button and a "LICENSE" button. It features a prominent green button labeled "Download the corpus for offline use". Below this is a text block about the corpus's size and growth, mentioning 13.6 billion words from 2010 to the present. It then discusses how the corpus differs from Google Trends, highlighting its structured nature and up-to-date data. Further down, it explains context-sensitive help, comparisons between dates and countries, and virtual corpora. Finally, it mentions the corpus's relation to other English corpora and its unparalleled insight into variation in English.

NOW Corpus (News on the Web) i file download

SEARCH FREQUENCY CONTEXT OVERVIEW

List Chart Collocates Compare KWIC

[POS]?

Find matching strings Reset

Search by date

Sections Texts/Virtual Sort/Limit Options

(HIDE HELP) + LICENSE

download Download the corpus for offline use

The NOW corpus ([News on the Web](#)) contains [13.6 billion words of data](#) from web-based newspapers and magazines from 2010 to the present time (the most recent day is [2021-10-13](#)). More importantly, the corpus grows by about 180-200 million words of data each month (from about 300,000 new articles), or about two billion words each year.

While other resources like [Google Trends](#) show you what people are *searching for*, the NOW Corpus is the only structured corpus that shows you what is actually *happening* in the language -- virtually right up to the present time. For example, see the [frequency of words](#) since 2010, as well as [new words and phrases](#) from the last few years.

Click on any of the links in the search form to the left (such as [List](#) or [Chart](#)) for context-sensitive help, and to see the range of queries that the corpus offers. You might pay special attention to the [comparisons](#) between dates and countries and [virtual corpora](#), which allow you to create personalized collections of texts based on (sub-)register, website, and even words in the web pages .

Finally, the corpus is related to many other [corpora of English](#) that we have created. These corpora were [formerly](#) known as the "BYU Corpora"), and they offer unparalleled insight into [variation in English](#).



7. Corpus tools and types of analysis

- Corpora are computer-readable collections of texts which enable linguistic analysis by means of special computer programs called **concordancers**
- The most popular concordancers are **WordSmith tools**, **Sketch Engine**, **MonoConc** and **AntConc**

<https://www.lexically.net/wordsmith/>

Windows software for finding word patterns

Published by [Lexical Analysis Software](#) and [Oxford University Press](#) since 1996



Concord

... for finding all instances of a word or phrase. [Video](#)

KeyWords

... helps find salient words in a text or set of texts. [Video](#)

WordList

... lists the words in your text(s) in alphabetical and frequency order. [Video](#)

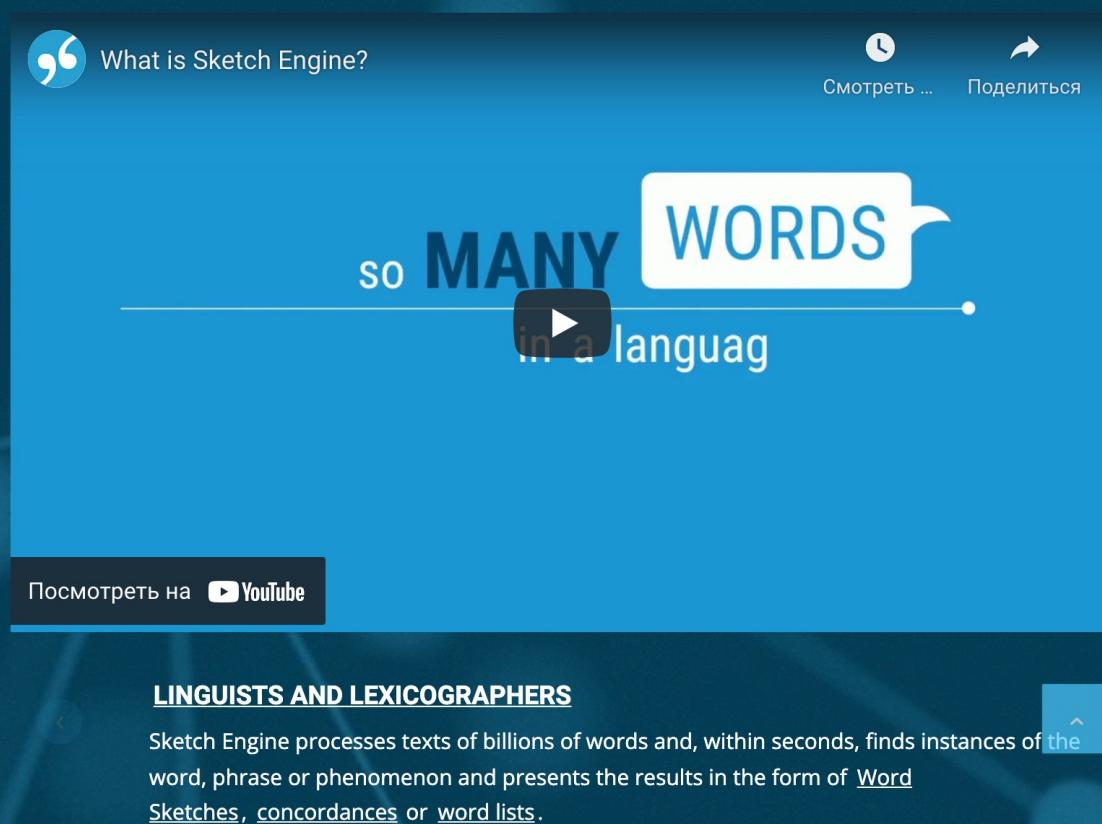
and a number of further [Utility tools](#)

What is Sketch Engine?

Sketch Engine is the ultimate tool to explore how language works. Its algorithms analyze authentic texts of billions of words (text corpora) to identify instantly what is typical in language and what is rare, unusual or emerging usage. It is also designed for text analysis or text mining applications.

Sketch Engine is used by linguists, lexicographers, translators, students and teachers. It is a first choice solution for publishers, universities, translation agencies and national language institutes throughout the world.

Sketch Engine contains 500 ready-to-use corpora in 90+ languages, each having a size of up to 60 billion words to provide a truly representative sample of language.



What is Sketch Engine?

Смотреть ... Поделиться

so MANY WORDS in a language

Посмотреть на  YouTube

LINGUISTS AND LEXICOGRAPHERS

Sketch Engine processes texts of billions of words and, within seconds, finds instances of the word, phrase or phenomenon and presents the results in the form of Word Sketches, concordances or word lists.

This site is primarily for users of MonoConc. (See [ParaConc](#) and [Collocate](#)) It is possible to order MonoConc Pro for immediate download from here, but for normal orders, visit athel.com or send email to info@athel.com.

monoconc.com

MonoconcEsy

A new version of the concordancer for use in computer labs and general language courses. It is similar to MonoConc Pro, however, it does not have the wordlist comparison feature (keywords) or the Advanced Sort menu. In addition, there are fewer options in Advanced Collocation. Price **Free for individuals** for non-commercial research. Site licence. \$290 15-users for 2 years.
Features of MonoconcEsy: Text Search and Part-of-Speech Tag Search and grep Search. Primary and secondary sorting of concordance lines. Frequency Wordlist. Collocate Frequency. Download from michaelbarlow.com/mcesy.zip

MonoconcEsy -- Chinese

The menus etc. are in Chinese. There are two versions, one for Simplified characters and one for Traditional

MonoConc Pro

Current version 2.2, released 2004
New features: Open/Save Workspace; Tag Search; Meta-Tags; Suppress/Display Word/Tag/Part-of-Speech Tag.
Price **\$85** Site licence. \$550 15-users for 2 years.
MonoConc Pro [pdf flyer](#)

To get the software immediately for individual use (**\$29**), click on the BUY NOW button. The transaction and link to the software is processed through e-junkie.com. You will be directed to Paypal where you can pay by credit card. You will then be given the link to download the zip file. You will also be sent the link via email.



Comments:

Corpus of Spoken Professional American-English

A 2-million word collection consisting of transcripts of American spoken in professional settings such as committee meetings, faculty meetings and White House press conferences. Tagged with part-of-speech tags. (CLAWS7 tagset)
Price **\$79**. For a basic search of the untagged version of the sibcorpora, you can use the following:

—Basic search: _____

Select subcorpus of CSPAE

Search term: _____

AntConc Homepage

Latest Release



AntConc

A freeware corpus analysis toolkit for concordancing and text analysis.

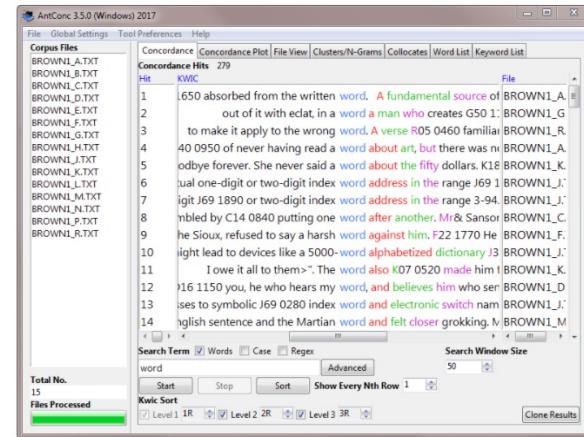
[[AntConc Homepage](#)] [[Screenshots](#)] [[Help](#)] [[License](#)]

Downloads:

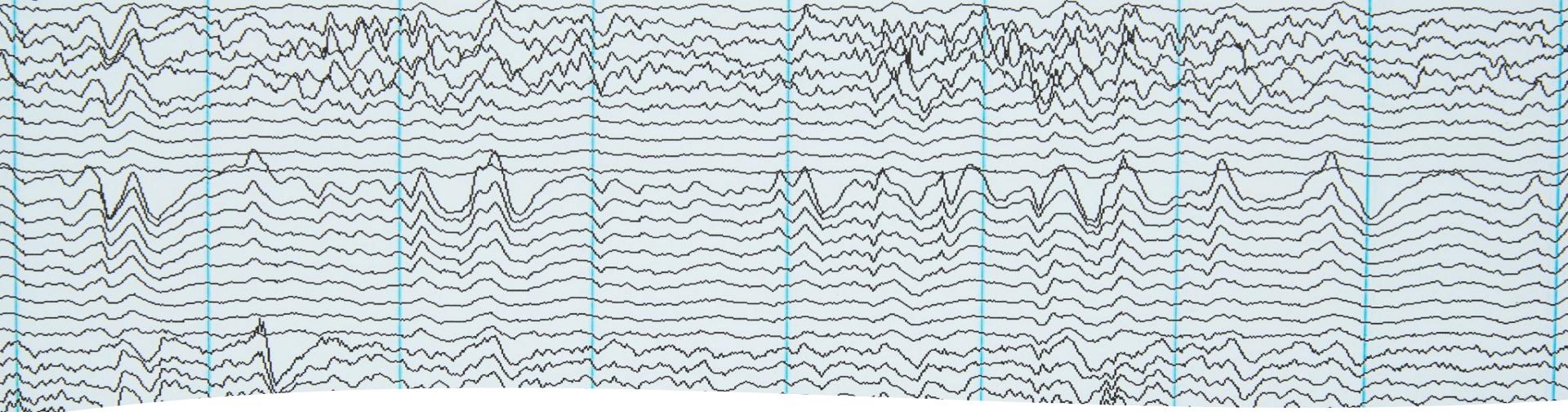
Official releases

- [Windows \(3.5.8\)](#)
- [Windows 64-bit \(3.5.9\)](#)
- [Macintosh OS X \(3.5.9\)](#)
- [Linux 64-bit \(3.5.9\)](#)

... .



- <https://www.laurenceanthony.net/software/antconc/>



7.1. Frequency analysis and concordancing

The most basic type of corpus analysis is checking the ***frequency of occurrence*** of a given word or a phrase

a search by means of a web-based interface

<https://www.english-corpora.org/>

Corpus (online access)	Download	# words	Dialect	Time period	Genre(s)
iWeb: The Intelligent Web-based Corpus		14 billion	6 countries	2017	Web
News on the Web (NOW)		13.6 billion+	20 countries	2010-yesterday	Web: News
Global Web-Based English (GloWbE)		1.9 billion	20 countries	2012-13	Web (incl blogs)
Wikipedia Corpus		1.9 billion	(Various)	2014	Wikipedia
Coronavirus Corpus		1.22 billion+	20 countries	Jan 2020-yesterday	Web: News
Corpus of Contemporary American English (COCA)		1.0 billion	American	1990-2019	Balanced
Corpus of Historical American English (COHA)		475 million	American	1820-2019	Balanced
The TV Corpus		325 million	6 countries	1950-2018	TV shows
The Movie Corpus		200 million	6 countries	1930-2018	Movies
Corpus of American Soap Operas		100 million	American	2001-2012	TV shows
Hansard Corpus		1.6 billion	British	1803-2005	Parliament
Early English Books Online		755 million	British	1470s-1690s	(Various)
Corpus of US Supreme Court Opinions		130 million	American	1790s-present	Legal opinions
TIME Magazine Corpus		100 million	American	1923-2006	Magazine
British National Corpus (BNC) *		100 million	British	1980s-1993	Balanced
Strathy Corpus (Canada)		50 million	Canadian	1970s-2000s	Balanced
CORE Corpus		50 million	6 countries	2014	Web



- use a **search box** located on the left-hand side of the interface
 - type in a word or a phrase that you want to explore (a **node**)
 - ‘**Chelyabinsk**’
 - word as a **lemma** = [chelyabinsk]

Information about the number of the occurrences of ‘Chelyabinsk’ in the whole corpus

 British National Corpus (BNC)  

SEARCH FREQUENCY CONTEXT

ON CLICK: [CONTEXT](#) [TRANSLATE \(??\)](#) [GOOGLE](#) [IMAGE](#) [PRON/VIDEO](#) [BOOK \(HELP\)](#)

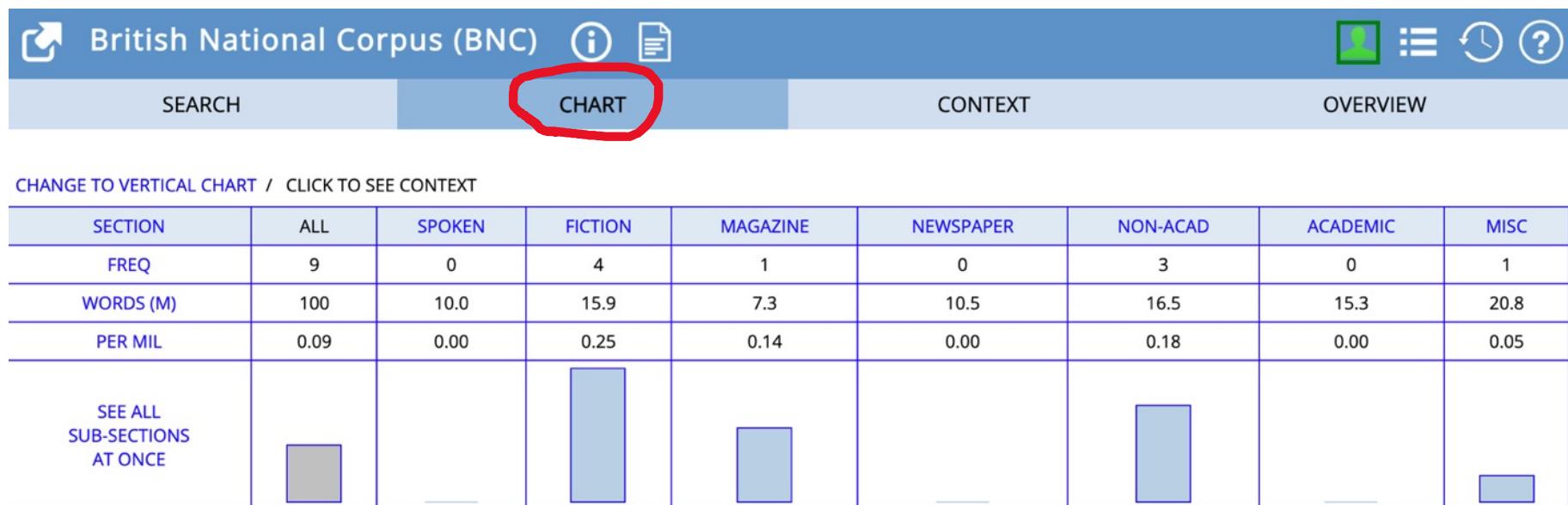
HELP		ALL FORMS (SAMPLE): 100 200 500	FREQ	
1	<input type="checkbox"/>	CHELYABINSK	9	

Lines of texts which demonstrate how the word is used in context: **concordances**



CLICK FOR MORE CONTEXT				EXPLORE NEW FEATURES 	 SAVE  TRANSLATE  ANALYZE
1	CLD	W_fict_prose	   Q	There are the Institutes and Design Laboratories of the Ministry of Medium Machine-Building in the Chelyabinsk region of the Ural mountains. In France, there are the	
2	CLD	W_fict_prose	   Q	.' The envelope was marked Personal and Confidential. As at Los Alamos and Chelyabinsk and Ripault and Lanzhou, the Atomic Weapons Establishment at Aldermaston	
3	CLD	W_fict_prose	   Q	work done here would be known to the scientists and engineers at Los Alamos and Chelyabinsk and Ripault and Lanzhou. But Los Alamos and Chelyabinsk and Ripaul	
4	CLD	W_fict_prose	   Q	engineers at Los Alamos and Chelyabinsk and Ripault and Lanzhou. But Los Alamos and Chelyabinsk and Ripault and Lanzhou and Aldermaston form the club with the	
5	CR8	W_pop_lore	   Q	In Russia itself, the nuclear industry that once supported whole cities, such as Chelyabinsk east of the Urals, and Tomsk and Krasnoyarsk in Siberia, is facing a	
6	HKT	W_non_ac_polit_law_edu	   Q	p. 36401 for such discoveries in January 1989 were discovered during 1989 near the towns of Chelyabinsk (80,000 corpses) and Donetsk and in Karelia and Byelorussia	
7	HKT	W_non_ac_polit_law_edu	   Q	of an accident in September 1957 at the top-secret Kyshtym nuclear weapons research facility near Chelyabinsk in the Ural mountains. On July 17 an account of the ac	
8	CP4	W_non_ac_tech_engin	   Q	service; the company also aims to set up new nodes in the cities of Chelyabinsk, Salavat and Sterlitomak in the near future. # AUTODESK RUSSIA STARTS AUTOCAD DI!	
9	J3D	W_misc	   Q	people are estimated to have been exposed to radioactive emissions from the Mayak plant near Chelyabinsk (which still produces plutonium and reprocesses waste) i	

All frequency values for the word ‘Chelyabinsk’ across the different portions of the corpus (Chart option)



Choose a different subcorpus

News on the Web



NOW Corpus (News on the Web)   

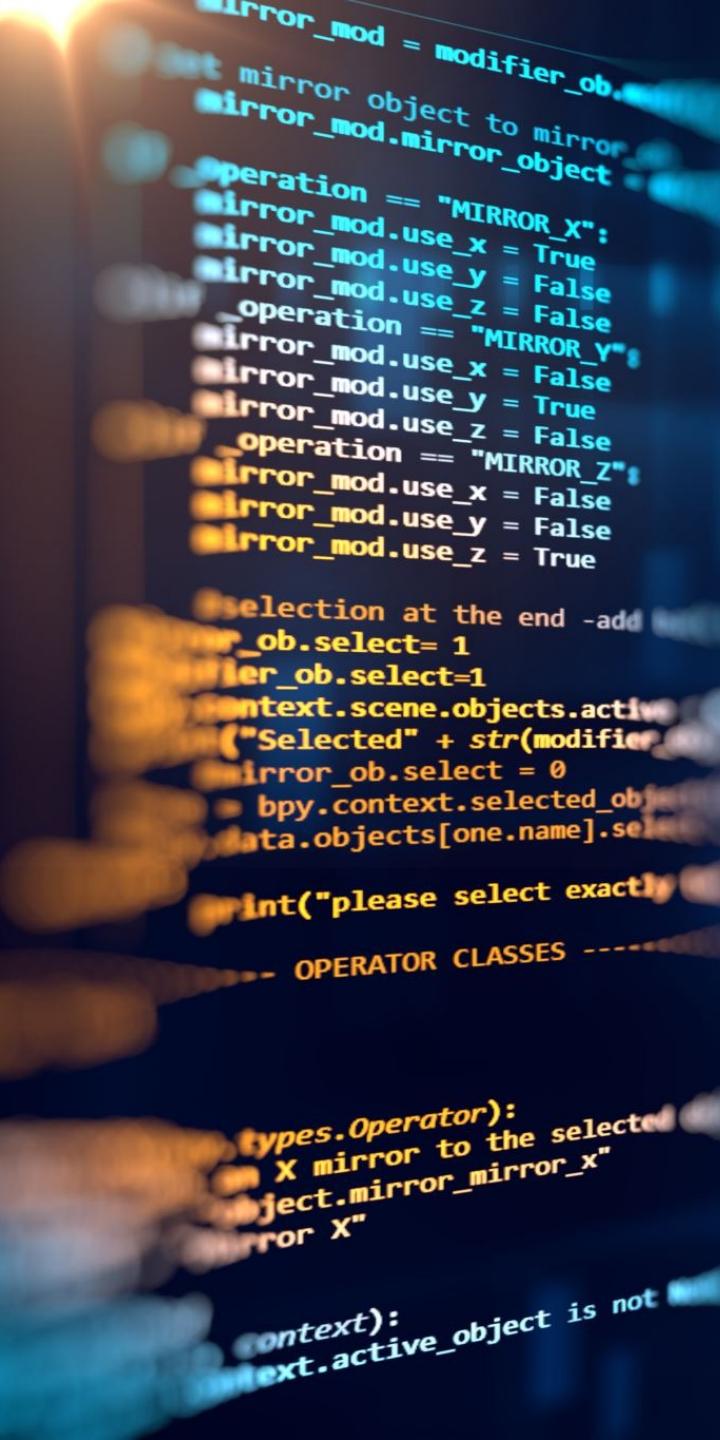
SEARCH FREQUENCY CONTEXT

ON CLICK:  CONTEXT  TRANSLATE (??)  GOOGLE  IMAGE  PRON/VIDEO  BOOK (HELP)

HELP		ALL FORMS (SAMPLE): 100 200 500	FREQ
1	<input type="checkbox"/>	CHELYABINSK	1914

Examples of concordances for the word ‘Chelyabinsk’

CLICK FOR MORE CONTEXT				EXPLORE NEW FEATURES	SAVE	TRANSLATE	ANALYZE
1	21-10-23 GB	Space.com		major city, but the blast occurred over a broad area outside the city of Chelyabinsk, resulting in damage and injuries but no fatalities). # To i			
2	21-10-23 GB	Space.com		# "An impactor the size of the 20-meter-wide space rock that broke up over Chelyabinsk, Russia... could be intercepted a mere 100 seconds			
3	21-10-23 GB	Space.com		large asteroid, like the 62-foot wide (19 m) meteor that exploded over Chelyabinsk, Russia in February 2013 with roughly the strength of 30			
4	21-10-23 GB	Space.com		objects can still pack a huge punch. # Part of the reason that the Chelyabinsk meteor was so destructive is that astronomers didn't see it co			
5	21-10-23 GB	Space.com		's success also hinges on scientists' ability to detect small near-Earth asteroids like the Chelyabinsk impactor before they enter the atmosph			
6	21-10-23 GB	Space.com		with diameters greater than 460 feet (140 m). However, as the Chelyabinsk incident showed, smaller objects can still pack a huge punch. # F			
7	21-10-21 US	Bulletin of the Atomic Scientists		# A remarkable Soviet-era monument lies on the outskirts of the formerly secret city of Chelyabinsk. It features a larger-than-life statue of Ig			
8	21-10-21 US	Bulletin of the Atomic Scientists		imagined. Receive Email Updates Monument to the Splitting of the Atom # Chelyabinsk City, Russia May 18, 1992 # A remarkable Soviet-era			
9	21-10-21 US	Bulletin of the Atomic Scientists		such as Alvin Weinberg's experimental molten salt reactor. Maids of Muslyumovo # Chelyabinsk Oblast, Southern Urals Region, Russia May			
10	21-10-17 US	New York Post		had he accepted his AHL assignment. Instead, the 21-year-old is back home in Chelyabinsk, Russia, The Post can confirm. # It appears the R			
11	21-10-15 US	YAHOO!Finance		gas pipeline project is seen on a pipe at the Chelyabinsk pipe rolling plant in Chelyabinsk, Russia # In this article: # Oops! # Something went			
12	21-10-15 US	YAHOO!Finance		of the Nord Stream 2 gas pipeline project is seen on a pipe at the Chelyabinsk pipe rolling plant in Chelyabinsk, Russia # In this article: # Oo			
13	21-10-15 AU	The Conversation		Earth. The impact hazard of asteroids is relatively well publicised, particularly after the Chelyabinsk meteor which exploded over a Russian t			
14	21-10-06 GB	Space.com		, " he said. " About 100,000 times more than the energy of the Chelyabinsk meteor and a million times more energy than the bombs droppe			
15	21-10-03 NG	gistmania.com		major cities in Russia's Urals region -- Chelyabinsk, Yekaterinburg and Tyumen. In Chelyabinsk, witnesses said the # Actress Foluke Daramol			
16	21-10-03 NG	gistmania.com		The meteor shower sparked panic in three major cities in Russia's Urals region -- Chelyabinsk, Yekaterinburg and Tyumen. In Chelyabinsk, w			
17	21-09-29 US	CNN		, 2021, aboard a helicopter surveying damage from wildfires in some areas of the Chelyabinsk region, Russia. # Moscow (CNN) Russian Eme			



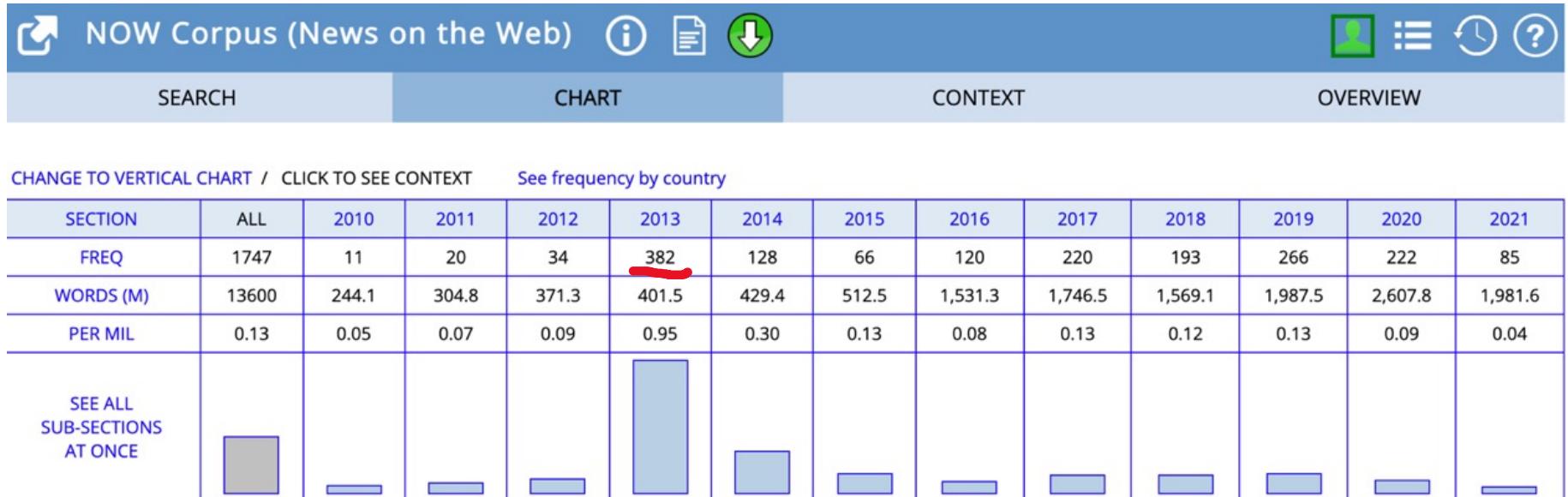
a standard format for displaying corpus data

Key Word In Context (KWIC)

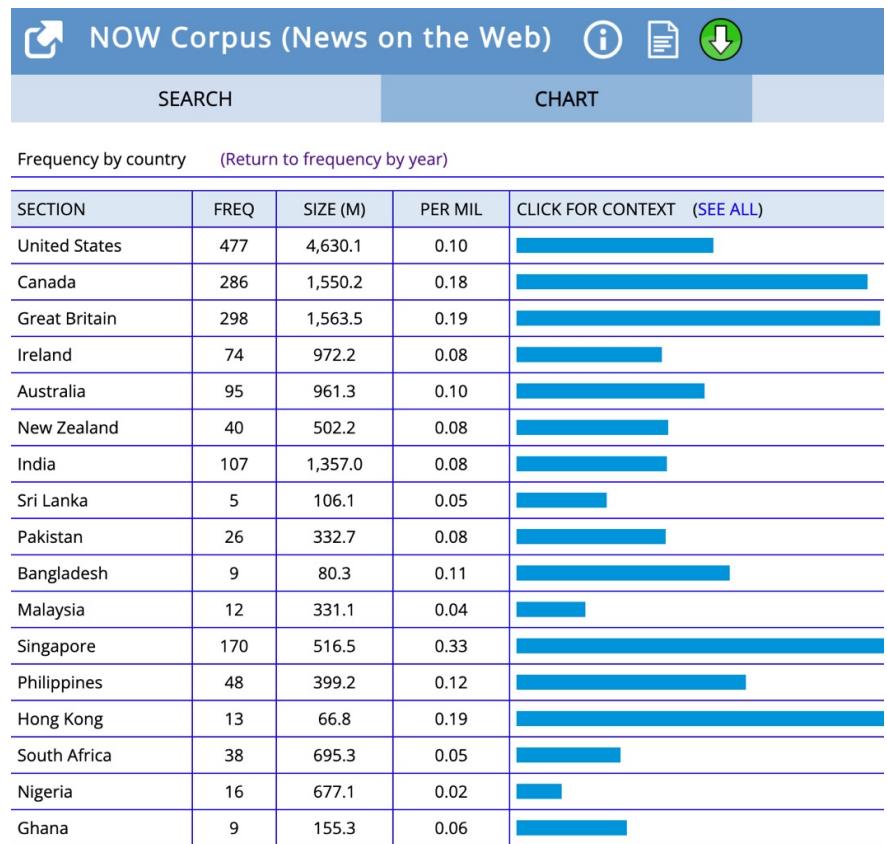
one can easily analyze the context of the node – all the words that precede and follow it

- By analyzing the immediate company of words, one can explore *patterns of co-occurrence* between words and study how words tend to form various kinds of *lexical, grammatical, lexicogrammatical combinations*

Frequency of words across different sections



Frequency of the word **‘Chelyabinsk’** *by country*



7.2. Wordlists

lists of words or phrases ranked according to their frequency or the number of their occurrences in a given corpus

wordlists are a powerful tool for making comparisons between corpora that represent different language uses

if we assume that the most frequent words are also the most useful ones, language teachers can use this information to decide which words should be addressed first where English is taught as a second/foreign language

COCA

Corpus of Contemporary American English

search for words by meaning ‘*industrial*’

Meaning	<input type="text" value="industrial"/> + <input checked="" type="checkbox"/> DEFINITION <input type="checkbox"/> SYNONYM <input type="checkbox"/> SPECIFIC <input checked="" type="checkbox"/> GENERAL
<p>● You can now search for words by meaning. For example, words with the following words in the definition: sugar, molecul*, magic*. You can also add a second word [-] for the dictionary entry, e.g. herb OR herbs (herb* would include the perhaps unwanted herbivore as well), computer AND device, cloud* NOT network. You can also search by synonym (noun: festival, disaster; adjective: harsh, kind; verb: groan, laugh), find more specific words (noun: machine, toy; verb: cry, walk) or more general words (frisbee, tequila; shriek, sashay) (both for just nouns/verbs), or combine these (e.g. walk, scare, screen, crystal).</p>	
Part of speech	<input checked="" type="checkbox"/> NOUN <input checked="" type="checkbox"/> VERB <input checked="" type="checkbox"/> ADJ <input checked="" type="checkbox"/> ADV <input type="checkbox"/> OTHER <input type="checkbox"/> ALL
Range	<input type="text"/> - <input type="text"/>
Pronunciation	Rhymes with <input type="text"/> Type <input type="text"/> EXACT
Syllables / stress	

FREQ	Word	PoS	Audio	Video	Image
373744	business	NOUN	🔊	▶	🖼️

enterprise and the people who constitute it 2 . the activity of providing goods and services involving financial and business concerns collectively 4 . the volume of business activity 5 . a rightful concern or responsibility

118323	plant	NOUN	🔊	▶	🖼️
--------	-------	------	---	---	----

industrial labor 2 . a living organism lacking the power of locomotion 3 . something planted secretly for discovery 4 . those acting is rehearsed but seems spontaneous to the audience

110976	concern	NOUN	🔊	▶	🖼️
--------	---------	------	---	---	----

ts you because it is important or affects you 2 . an anxious feeling 3 . a commercial or industrial enterprise and the one that causes anxiety 5 . a feeling of sympathy for someone or something

49867	works	NOUN	🔊	▶	🖼️
-------	-------	------	---	---	----

industrial labor 2 . performance of moral or religious acts 3 . everything available 4 . the internal mechanism of

35394	tech	NOUN	🔊	▶	🖼️
-------	------	------	---	---	----

anical and industrial arts and the applied sciences

27047	plant	VERB	🔊	▶	🖼️
-------	-------	------	---	---	----

edlings) into the ground 2 . fix or set securely or deeply 3 . set up or lay the groundwork for 4 . put firmly in the mind in order to secretly observe or deceive

19955	concern	VERB	🔊	▶	🖼️
-------	---------	------	---	---	----

mind of

15934	technological	ADJ	🔊	▶	🖼️
-------	---------------	-----	---	---	----

7.3. Word combinations and *n-gram analysis / cluster analysis*

chunks

n-grams

lexical bundles

“

Words tend to co-occur and form collocations, colligations and other examples of word combinations

“ ” “

N-gram is a technical term used to denote word combinations which consist of two or more words that repeatedly occur consecutively in a corpus

Corpus software AntConc

<https://www.laurenceanthony.net/software/antconc/>



AntConc

A freeware corpus analysis toolkit for concordancing and text analysis.

[[AntConc Homepage](#)] [[Screenshots](#)] [[Help](#)] [[License](#)]

Downloads:

- [Windows \(3.5.8\)](#)
- [Windows 64-bit \(3.5.9\)](#)
- [Macintosh OS X \(3.5.9\)](#)
- [Linux 64-bit \(3.5.9\)](#)
- [Older versions](#)
- [Win 10 \(4.0.0 RC 2.0\)](#)
- [MacOS 10 Catalina \(4.0.0 RC 2.0\)](#)
- [MacOS 11 Big Sur \(Intel+Silicon\) \(4.0.0 RC 2.0\)](#)

PayPal Donations and Patreon Supporters:

Click one of the following if you want to make a small donation to support the future development of this tool.

[Support this tool](#)

[BECOME A PATRON](#)

A screenshot of the AntConc 3.5.0 software interface. The window title is "AntConc 3.5.0 (Windows) 2017". The menu bar includes File, Global Settings, Test Preferences, and Help. The main menu is "Concordance". The sub-menu "Concordance Hits" shows 279 hits. The main pane displays a list of 15 hits from a corpus file named "BROWN1_KWIC". The hits are numbered 1 through 15 and show examples of words from the Brown corpus. The interface includes various search and filter options at the bottom.



7.4. Keynes analysis and keywords

- **Keyword** – a word which occurs with unusual frequency in a given text.
- Such words are useful because they provide information about the **keyness** or specificity of a given corpus in terms of what it is about.

Keywords tool on Lextutor

<http://www.lextutor.ca/key/>

KeyWords Extractor v. 2.2

The keywords of a text are more frequent in the text than in a reference corpus.

Input mode A: Type or paste small text (<50,000 words) below and click *Submit_window*

Title: Untitled

INSTRUCTIONS: Type or paste your text here and click the SUBMIT_window button,
or one of the sample texts below. Keyword analysis will...

TEXT SET-UP

General: Include an empty space after every comma or full stop.

Research: Deal with spelling errors and proper nouns.

SIZE LIMITS: Web input is OK up to 50,000 words, maybe more - to be safe use UPLOAD method below for larger files (up to 1000 kb or 1 MB as of Dec 2005; must be ~.txt; send in straight from your own hard drive). Text is NOT stored on server

5000+ Wd Samples: [Dracula](#) | [Love Story](#) | [Mutiny - Bounty](#) | [Jungle Book](#) | [Speckled Band](#) |

Reference corpus (for either input mode)

Example .txt format files

Trump inaugural address

[Home](#) > [Keywords Input \(or Back to preserve input\)](#) > [Keywords Output](#) - POTENTIAL KEYWORDS IN trump_ inaugural_address.txt (1,458 words)

Keywords are the words that are far more frequent in your text, as a proportion of its size, than they are in a reference corpus (here, the 10-million word mixed written-spoken, US-UK, developed by Paul Nation as basis for the first 2k of the BNC-Coca lists - see it -> [bnc_coca_fams_speechwrite_US_UK_per10mill](#)).

The number preceding each word in the output below is the number of times more frequent this word is in **your text** than it is in the corpus **bnc_coca_fams_speechwrite_US_UK_per10mill**. For example, the first item in the output **20576.00 politic** means that **politic** has 1 natural occurrences in 10,000,000 words, but 3 occurrences in your 1,458-word text. This would work out to $(3/1458) \times 10,000,000 = 20,576$ occurrences if your text were the same size as the corpus. The word is thus $20,576 / 1 = 20576.00$ times more frequent in your text than it is in the reference corpus. This probably means the word plays an important (or 'key') role in your text.

[Read more...](#)

The keyword list below contains all the words in your text that are at least **25 times** more numerous in your text than in the reference corpus (the "keyness factor."). The greater the keyness factor, the more 'key' a word is likely to be to your input text.

Paste raw (numberless, unsorted) output to other routines for comparison or alternative takes on output (Freq, VP,
Tex_Lex_Compare) [Get Raw List](#)

- (1) 20576.00 politic
- (2) 160.57 america
- (3) 91.04 prosper
- (4) 65.95 wealth
- (5) 61.79 factory
- (6) 56.37 border
- (7) 40.83 bless
- (8) 34.58 celebrate
- (9) 33.62 dream
- (10) 32.16 citizen
- (11) 29.29 protect

At keyness cut-off of 25, there are 11 keywords from a total of 1,458 words, for a keyword ratio of 7.54 per 1,000 words

WARNING: Short texts (<5,000 words) are not suitable for keyword analysis. This text is only 1,458 words

lists of keywords might be a *starting point*
for a qualitative analysis

concordance lines from the **inaugural address** Corpus ↗

investigate how the words '**america**'
and '**prosper**' are used in specific
contexts

