

A linguist's toolkit for anybody searching for authentic language material

Language corpora

Language corpora are in the broadest sense defined as collections of texts (cf. Kennedy 1998). They are compiled for linguistic analyses but are also a valuable source for finding authentic language material. Different corpora include different kinds of texts (ranging from transcriptions of audio-recorded informal conversations to newspaper articles or even scripts of soap operas). Corpora are electronically searchable for words and word classes but (unfortunately) not for pragmatic phenomena (such as speech acts, politeness strategies, etc.). While you can search a corpus for the word *please* (and the search engine will show you all the hits), you will not be able to locate all requests in the corpus that do not include this lexical item. Still, corpora are a valuable source; especially as some come with audio files!

Here's a selection of corpora which are freely available on the internet:

[Santa Barbara Corpus of Spoken American English](#)

Size	249,000 words
Variety	American English
Medium	Spoken English
Genres	Naturally occurring mostly informal face-to-face conversations
Sampling period	1990s
Audio files?	Yes! (freely downloadable: very nice tool for including authentic audio material in the classroom)

[Michigan Corpus of Spoken Academic English \(MICASE\)](#)

Size	1.8 million words
Variety	American English
Medium	Spoken English
Genres	Academic English (Lectures, seminars, counselling sessions, etc.)
Sampling period	1997-2001

[Vienna-Oxford International Corpus of English \(VOICE\)](#)

Size	1 million words
Variety	English spoken as a lingua franca by speakers of more than 50 L1s
Medium	Spoken English
Genres	Naturally occurring, non-scripted face-to-face interactions
Sampling period	2001-2007
Comment	Collection of naturally occurring interactions of non-native speakers of English (English used as a lingua franca)

[Time Magazine Corpus](#)

Size	More than 100 million words
Variety	American English
Medium	Written English
Genres	Articles from the Time Magazine
Sampling period	1923 – 2006 (all issues of Time Magazine)

[The British National Corpus \(BNC\) Online](#)

Size	100 million words
Variety	British English
Medium	90% written and 10% spoken language
Genres	Spoken, fiction, magazine, newspaper, non-academic, academic
Sampling period	1970s - 1993

Links: Language variation/ diversity

If you are interested in specific varieties of English and looking for authentic language material, there is a multitude of helpful and very good sites online. Here's a short annotated selection of links where you can stream or even download authentic language material to use in the classroom (or anywhere else, really). All of the sites have been developed either by linguists or in collaboration with linguists (i.e. the materials are much more reliable than what you find on most other sites).

[Do you speak American?](#)

Do you speak American? is a highly entertaining and very informative PBS production that explores features of American English and tries to answer questions like "Why is the English spoken by Maine lobstermen so different from that spoken by cowboys in Texas?" or "Does Spanish pose a threat to English as the dominant language in America?" for a non-linguist audience. Journalist Robert MacNeil travels cross-country to answer these questions and examine the dynamics of American English - a language rich with regional variety, strong global impact and cultural controversy.

Highly recommended also for the advanced EFL classroom! The website offers a lot of ready-made materials for teachers.

The DVDs are available in the Mediathek (BIS Oldenburg).

[Sounds familiar? Accents and dialects in the UK](#)

Sounds familiar? is a project by the British Library that maps sound recordings of speakers from all over the UK. The site includes many recordings, transcripts of them and additional information on many topics relating to accents and dialects in the UK. It even contains a glossary, in case you are not sure what exactly double negation means. On top of the easy to understand definition, the glossary also links you directly to those sound files in which the phenomenon you were looking for occurs.

[BBC Voices: Accents and dialects in the UK](#)

BBC Voices is a BBC project that collects and displays recordings of speakers from all over the UK. If you want to know whether friends on the Shetland Islands sound in contrast to a group of neighbours in Devon, BBC voices is the place to go!

[UK Sound Map](#)

The British Library project *UK Sound Map* is similar to BBC Voices in aims and scope. *UK Sound Map* is an interactive project that collects and displays sound recordings of speakers from all over the UK.

[International Dialects of English Archive \(IDEA\)](#)

The International Dialect of English Archive is an online archive of primary source dialect and accent recordings for the performing arts. It includes information about and sound recording of dialects or varieties of English spoken all over the world.

[Atlas of North American English \(ANAE\) and the TELSUR Project](#)

The Telsur Project is a survey of linguistic changes in progress in North American English. It is the creator of the Atlas of North American English (ANAE).

"The *Atlas of North American English* provides the first overall view of the pronunciation and vowel systems of the dialects of the U.S. and Canada. The *Atlas* re-defines the regional dialects of American English on the basis of sound changes active in the 1990s and draws new boundaries reflecting those changes. It is based on a telephone survey of 762 local speakers, representing all the urbanized areas of North America. It has been developed by Bill Labov, one of the leading sociolinguists of the world, together with his colleagues Sharon Ash and Charles Boberg." (Publisher's information)

An [online version](#) of the atlas is available for students of Oldenburg University. Please note that you might only be able to access the online version if you are connected to the University network.