



ELTE

FACULTY OF
HUMANITIES

English Historical linguistics and the use of corpora

Alexandra FODOR

Linguistic Theories lecture course

24 November 2025

OUTLINE

- I. Changes in the history of English
- II. Introduction to Corpus Linguistics
 - a) Key concepts
 - b) Corpora
 - c) Applications
 - d) Why use corpora?
 - e) What can we do with corpora?
 - f) Corpus linguistics – what is it and what NOT?
 - g) Tools
 - h) Sample task

Lord's Prayer



Old English (11th c.)

Matthew 6.9 (WSCp, 11th c.) (Corpus Christi College MS 140, ed. Liuzza (1994))

Fæder ure þu þe eart on heofonum;
Si þin nama gehalgod
to becume þin rice
gewurpe ðin willa
on eorðan swa swa on heofonum.
urne gedæghwamlican hlaf syle us todæg
and forgyf us ure gyltas
swa swa we forgyfað urum gyltendum
and ne gelæd þu us on costnunge
ac alys us of yfele sopllice.

Present-day English

The Alba House New Testament (1970, tr. Condon)

Our Father in Heaven, let your holy name be known, let your kingdom come, and your will be done, on earth as in heaven. Give us today the bread that we need, and forgive us our wrongs, as we forgive those who have done wrong to us. Do not lead us into trial, but save us from evil.

Some peculiarities of Old English morphology

Old English morphology is **synthetic, inflectional, and structurally rich**, with:

- multiple cases
- agreement marking
- complex verb classes
- abundant prefixes and suffixes
- grammatical gender
- numerous declensional patterns

Compared with Present-Day English, OE expresses much more grammatical information through **morphological endings rather than word order**.

Some peculiarities of Old English syntax

- ✓ Freer (but not random!) word order supported by rich case morphology
- ✓ Verb-second tendencies in main clauses
- ✓ Multiple negation (negative concord)
- ✓ pre-verbal negation
- ✓ Flexible positioning of adjectives, possessors, and pronouns
- ✓ High use of subordination with *þe* and *þæt*

Some peculiarities of Old English semantics

Old English words typically cover **wider semantic ranges** than their Present-Day English equivalents.

- *yfel* → harm, misfortune, wickedness, evil (PDE: mainly moral evil)
- *costnung* → testing, trial, temptation
- *hlāf* → bread, food, sustenance, “what keeps life going”
- *heofonum* (heavens) = literal sky + divine realm
- *Fæder* conveys familial authority + divine sovereignty
- *forgyfan* originally “give fully, give up, release” → later “forgive”
- *willa* = intention, desire, plan, decree → stronger than PDE “will”
- *nama* = name + essence/identity
- *gylt* = crime, offence, guilt, sin (legal + moral + spiritual)

Some peculiarities of Old English sounds and spelling

Old English spelling is much more sound-based than Present-Day English, but it is not perfectly consistent.

Old English represents a much more transparent, phonologically faithful spelling system than PdE, yet one shaped by dialect, scribal habits, and phonological complexity (palatalisation, vowel length, diphthongs, consonant length).

Before front vowels, consonants like /k/ and /g/ undergo palatalisation

Vowel length is phonemic, and spelling often (but not always) reflects it: *god* (/god/) “god” vs *gōd* (/go:d/) “good”.

Letters in OE but not in PdE:

- Thorn <þ> and eth <ð> represent dental fricatives, but they do not distinguish voiced vs voiceless.
- Ash < æ >

Introduction to Corpus linguistics

Key concepts

- **Corpus**
- **Types**
- **Representativeness**
- **Annotation**

What is a corpus?

- Corpus (pl. corpora) = ‘body’
- Collection of written text or transcribed speech
- Electronically stored
- Usually but not necessarily purposefully collected
- Sources of data for language analysis
- Availability: online or software

What is a corpus?

- A corpus is a principled collection of authentic texts stored electronically that can be used to discover information about language that may not have been noticed through intuition alone.
- There are various types of corpora – depending on the actual need for research.

Corpora

- [Corpus Resource Database \(CoRD\)](#)
- [English Corpora: most widely used online corpora. Billions of words of data: free online access](#)

What corpus types exist?

Monolingual:

General(ised)/Reference

Specialised

Learner

Pedagogic

Historic /diachronic

Multilingual:

Parallel

Comparable

What corpus types exist?

Monolingual:

General(ised) / Reference: various text types, written and spoken, relatively large (more than 10 million words), e.g. BNC, ANC, COCA

Specialised: contains texts from a certain genre and aims to be representative of the language of this type, e.g. Corpus of Early Modern English Tracts, MICASE, MICUSP (Michigan Corpus of Upper-level Student Papers), Air Traffic Control Speech Corpus.

What corpus types exist?

Learner: a collection of texts from learners of a particular language.

E.g. ICLE (International Corpus of Learner English)

Pedagogic: contains language used in classroom settings; include academic textbooks, transcripts of spoken interaction in classroom. It can be used to ensure students are learning useful language, to examine teacher-student dynamics, or tool for teacher development.

Historic /diachronic: collection of texts from different periods of a language. It can help tracing the development and language change over time. E.g. HC, PPC(E)ME

What corpus types exist?

Multilingual:

Parallel: consists of at least 2 subcorpora compiled from different languages, e.g. EU texts. Can be used by translators, for instance.

Comparable: contains components in two or more languages that have been collected using the same sampling method, e.g. the *same proportions* of the texts of the *same genres* in the *same domains* in a range of *different languages* in the *same sampling period*. E.g. [LOB corpus sampling frame](#) for the [Lancaster Corpus of Mandarin Chinese](#) (McEnery *et al.* 2003)

What does **Representativeness** mean?

According to Leech (1991: 27), a corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to the said language variety.

Biber (1993: 243) : 'Representativeness refers to the extent to which a sample includes the full range of variability in a population.'

A corpus is essentially a sample of a language or language variety (i.e. population). Sampling is entailed in the compilation of virtually any corpus of a living language. In this respect, the representativeness of most corpora is to a great extent determined by two factors: the range of genres included in a corpus and how the text chunks for each genre are selected

What does **annotation** mean?

Annotation refers to the addition of metalinguistic information to corpus texts:

Corpus annotation adds value to a corpus in that it considerably extends the range of research questions that a corpus can readily address.

Types of annotation

- **Part-of-Speech (POS) Tagging**: The most common type, where a label indicating the word class (e.g., noun, verb, adjective) is assigned to each word. For example, "talk" in "I heard John's talk" would be tagged as a noun.
- **Lemmatisation**: The process of grouping together the inflected forms of a word so they can be analyzed as a single item, or lemma (dictionary headword). For example, *going*, *goes* would be tagged with the lemma **GO**.
- **Syntactic Annotation (Parsing/Treebanking)**: Involves adding information about the grammatical structure of sentences, such as how words form phrases and clauses.
- **Semantic Annotation**: Adding information about the *meaning* of words or phrases, such as distinguishing between "cricket" (sport) and "cricket" (insect).

Types of annotation

- **Pragmatic and Discourse Annotation:** Adding information about the function of an utterance (e.g., as a command, question, or acknowledgment) or linking pronouns to their antecedents (e.g., "I'll saddle the horses and bring *them* round").
- **Phonetic and Prosodic Annotation:** For spoken corpora, this involves adding details about pronunciation, stress, intonation, and pauses.
- **Error Tagging:** Used in learner corpora to indicate the *type and location* of the errors made by second-language learners; this can be useful for language pedagogy research.

Why use corpora?

- Corpora allow linguists to access quantitative information about language, which can often be used to support qualitative analysis.
- Insights into language gained from corpus analysis are often generalisable in a way that insights gained from the qualitative analysis of small samples of data are not.
- Using corpus data forces us to acknowledge how language is really used (which is often different from how we think it is used).

Applications

- **Historical linguistics** → internal developments in language, e.g. grammaticalisation (would from main verb to auxiliary)
- **Syntax** → e.g. major innovations in the development of grammar of spoken language.
- **Semantics** → Sinclair (1991, 2004) collocation and phraseology
- **Dictionary writing** → Collins Birmingham University International Language Database = COBUILD
 - An electronic corpus of cca. 2.5 billion words of contemporary British English
 - COBUILD dictionary: 1987; based on evidence from the corpus, and not on intuitions or created instances.

Applications

- **Sociolinguistics** → Gender, dialect, identity
- **Discourse Analysis** → Media, politics, ideology
- **EAP** → Academic writing, learner corpora
- **Stylistics** → Literary analysis, authorial voice

What can you do with a corpus?

- Frequency analysis
- Keyword analysis
- Concordance analysis
- Collocation analysis

What can you do with a corpus?

- Frequency analysis:
 - how often a particular word is used in the texts / text types
 - Investigate lexical change across time or differences between texts
- Keyword analysis: to discover the statistically significant words in a corpus.
- Concordance analysis: to investigate a keyword's meaning and its grammar in context, and its possible variation.
- Collocation analysis: to investigate the typical collocates of the given keyword. Collocates are words that have a tendency to co-occur with the keyword.



English-Corpora.org

[corpora](#) [PDF guides](#) [videos](#) [related resources](#) [users](#) [my account](#) [upgrade](#) [help](#)

Overview: [brief](#) | [detailed](#) 

Now available: AI/LLM insights integrated into corpus results

These are the most [widely used](#) online corpora, and they serve [many different purposes](#) for teachers and [researchers](#) at [universities](#) throughout the world. In addition, the corpus data (e.g. [full-text](#), [word frequency](#)) has been employed by a [wide range of companies](#) in many different fields, especially technology and [language learning](#).

The links below are for the free online interface. You can also purchase and download  the corpora for use on your own computer.

Corpus	Download	# words	Dialect	Time period	Genre(s)
News on the Web (NOW)		23.5 billion+	20 countries	2010- yesterday	Web: News
iWeb: The Intelligent Web-based Corpus		14 billion	6 countries	2017	Web
Global Web-Based English (GloWbE)		1.9 billion	20 countries	2012-2013	Web (incl blogs)
Wikipedia Corpus		1.9 billion	(Various)	2014	Wikipedia
Coronavirus Corpus		1.5 billion	20 countries	2020-2023	Web: News



SEARCH

WORD

CONTEXT

ACCOUNT

 List Chart **Word** Browse + [POS] ?[See detailed info for word](#)[Reset](#)AI: search by **concepts**, translations, rephrasing (HIDE HELP)**NO LICENSE**

You can find a wealth of information for the top 60,000 words in the corpus. As the following examples with *bread* show, you can see:

- an [overview](#) of all of the information below
- related [topics](#) (words that co-occur *anywhere* on the web page)
- [collocates](#) (automatically grouped by part of speech)
- [clusters](#) (the most frequent 2, 3, and 4 word strings)
- a resortable [Keyword in Context](#) (concordance) display
- [related words](#) (synonyms and WordNet entries), and
- [websites](#) use that word the most (can use these to create Virtual Corpora).

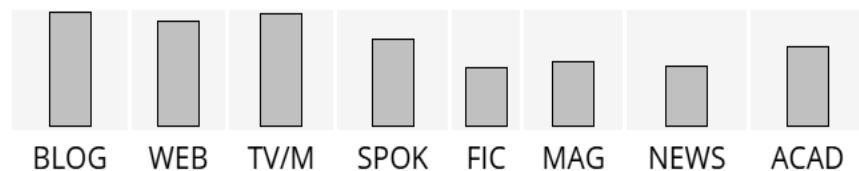
should

(VERB)  

#110 

Freq: 920,908 Range: 236,459 texts (48.7%)

Forms: should (920,908)



1. expresses an emotional, practical, or other reason for doing something 2. be expected to 3. be logically necessary

D M O C G **E**

  YouGlish PlayPhrase Yarn

 Translate: choose language

SYNONYMS (▶ CONCEPT)

NEW: DEFIN +SPEC +GENL

To the left is frequency information for [should \(v\)](#) .

It shows the total **frequency** of the word (lemma) in COCA, as well as the **range** -- the number of texts (from the 485,202 texts in the corpus) in which the word occurs, and the corresponding percentage of all texts that have the word. It also shows the frequency of each **form** of [should \(v\)](#) .

TOPICS (more)

[teacher](#)

COLLOCATES (more)

NOUN [us](#), [obama](#), [america](#), [israel](#), [clinton](#), [washington](#), [romney](#), [caution](#)

VERB [construe](#), [heed](#), [impeach](#), [commend](#), [disqualify](#), [recuse](#)

ADJ [ashamed](#), [accountable](#), [cognizant](#)

ADV [ideally](#), [sparingly](#)

should

(VERB)



#110



BLOG



WEB



TV/M



SPOK



FIC



MAG



NEWS



ACAD

1. expresses an emotional, practical, or other reason for doing something 2. be expected to 3. be logically necessary

D M O C G **E**



YouGlish PlayPhrase Yarn



Translate: choose language

See in iWeb



Collocates

Clusters

Topics

Texts

KWIC

TOPICS (more)

teacher

COLLOCATES (more)

NOUN

us, obama, america, israel, clinton, washington, romney, caution

VERB

construe, heed, impeach, commend, disqualify, recuse

ADJ

ashamed, accountable, cognizant

ADV

ideally, sparingly

	WEBSITE		SORT	SORT	SORT
1	BLOG:2012:freethoughtblogs.com	that enforced masculinity hurts men -- it does . But you	should	a) pick	your arguments more carefully , this one is n't
2	ACAD:2002:SportBehavior	low , medium , or high . Hence , different athletes	should	achieve their best	when just inside their own optimal zone , no
3	MAG:1999:PsychToday	women are raised to believe different things about the way they	should	actually handle money	. Despite many social changes , men are
4	WEB:2012:classics.mit.edu	those , (After the manner of the Giants) ,	should	all Pay the	huge penalties for monstrous crime , Who by their
5	ACAD:1999:GeographRev	simple words , ones in wide and varied use . They	should	allow different interpretations	, so that the public can craft
6	BLOG:2012:...yourlivingonline....	and fillers . Your opening statement is very important -- it	should	allow readers to	want to read the article to the end .
7	WEB:2012:bleacherreport.com	get the ball on the green to make par , he	should	also be a	factor at Augusta and the PGA . How
8	ACAD:1999:SportBehavior	were more likely to win . # (n2) It	should	also be noted	that two other variables were considered for
9	ACAD:1994:AmerScholar	serious books -- or that , as a woman , she	should	also feel daunted	by either prospect . In the de Mille home
10	TV:2005:Alias	MacLaine . Good to know you . I think . I	should	also mention	hear confession every week day from 4:00 to 6:00
11	WEB:2012:...ew.toptenreviews....	engine . For a more personalized web experience , the browser	should	also provide the	ability to save bookmarks , set parental
12	SPOK:2005:Fox_Saturday	did they overact anyhow ? GABLER : Well , the media	should	always be extremely	cautious in situations like this , in the
13	MAG:2009:Cosmopolitan	answer , leave a brief VM . RULE : A guy	should	always pay when	you 're on a first date ... UNLESS :
14	FIC:2007:Analog	generally southerly . When he could , he supposed , he	should	angle back and	pick up US 95 again . Something bright caught
15	WEB:2012:hellogiggles.com	he committed a crime . If we use that logic ,	should	any person who	does this be punished by not earning a living
16	TV:2014:Grey's Anatomy	fellowship , if you hear of anybody good . Maybe you	should	apply . Yeah	, then you could learn to do something that

should

(VERB) ★ ↻

#110 +



1. expresses an emotional, practical, or other reason for doing something 2. be expected to 3. be logically necessary

D M O C G E



YouGlish PlayPhrase Yarn



Translate: choose language

See in iWeb



Collocates

Clusters

Topics

Texts

KWIC

TOPICS (more)

teacher

COLLOCATES (more)

NOUN

us, obama, america, israel, clinton, washington, romney, caution

VERB

construe, heed, impeach, commend, disqualify, recuse

ADJ

ashamed, accountable, cognizant

ADV

ideally, sparingly



ELTE

FACULTY OF
HUMANITIES

COLLOCATES

SHOULD

VERB

Advanced options

🏠

🔄

Collocates

Clusters

Topics

KWIC

Use AI to group

+ NOUN	NEW WORD			?
1250	2.14	priority		
945	2.00	consideration		
667	3.24	caution		
294	2.27	correspondence		
244	2.36	policymaker		
191	2.54	clinician		
132	2.15	retrospect		
111	2.27	hindsight		
109	2.43	wake-up		
106	2.79	precedence		
82	3.17	reprint		
70	2.70	no-brainer		
34	2.01	heed		

+ ADJ	NEW WORD			?
11870	2.52	able		
1630	4.20	ashamed		
812	2.32	grateful		
664	3.00	accountable		
344	2.45	thankful		
326	2.36	cautious		
298	2.59	wary		
166	2.02	outraged		
143	2.56	mindful		
134	2.50	exempt		
114	2.47	vigilant		
52	2.95	cognizant		
51	2.14	off-limits		
35	2.04	payable		

+ VERB	NEW WORD			?
9634	2.41	consider		
4989	2.27	note		
967	2.20	apologize		
954	2.25	interpret		
939	2.74	punish		
782	2.17	ban		
781	2.46	resign		
734	2.54	consult		
664	2.83	strive		
477	2.98	underestimate		
440	2.35	prosecute		
435	2.35	dictate		
401	2.06	forbid		
341	3.69	construe		
320	2.88	reconsider		

+ ADV	NEW WORD			?
27977	2.18	why		
22681	3.03	maybe		
8127	2.02	probably		
5988	2.08	perhaps		
2722	2.03	therefore		
758	4.17	ideally		
450	2.19	sometime		
219	2.50	theoretically		
154	2.35	cautiously		
143	2.38	logically		
98	2.24	eg		
84	2.33	rightfully		
68	3.25	sparingly		
53	2.32	criminally		
38	2.80	snugly		

What is corpus linguistics?

What is corpus linguistics?

An approach to study the actual language in use through corpora.

John Sinclair (1990): a word in and of itself does not carry meaning, but that meaning is often made through several words in a sequence → backbone of CL.

Corpus Linguistics intends to answer two fundamental research questions:

- ✓ What particular patterns are associated with lexical or grammatical features?
- ✓ How do these patterns differ within varieties and registers?

What is corpus linguistics?

- **methodology** for studying large quantities of language data using computer software (Hardie and McEnery 2010).
- a **sub-discipline of linguistics** in its own right, concerned with explaining the relationship between meaning and structure in language (Tognini-Bonelli 2001:1).

Linguistics in general describe or explain a certain aspect of language use. CL, however is not restricted to a particular aspect of language; rather it can be employed to explore almost any area of linguistic research.

What is corpus linguistics NOT?

Corpus linguistics has its limits! →

Corpus linguistics is not

- able to provide negative evidence
- able to explain WHY?
- able to provide all possible language at a time

The history of Corpus Linguistics

Corpus linguistics: past and present

- Earliest forms at the beginning of the 20th century – shoeboxes filled with paper slips 😊 → Jespersen (1909-49), Fries (1952) used ,paper-based corpora' to study grammar
- Chomsky (1957) Syntactic Structures – with invented examples → corpora heavily criticised, thus abandoned
- Development of technology: ,real' start around the 1960s with BROWN corpus (→ ctd)
- CL first appeared in early 1980s

Corpus linguistics: past and present

Surveying English

1st project of CL: Survey of English Usage → 1 million words, written&spoken English, 200 text samples, 5000 words/text

By Randolph Quirk (1959), University College of London

Data on paper, indexed on file cards, manually annotated (for prosodic and paralinguistic features).

Corpus linguistics: past and present

- The **BROWN** corpus, 1961

Brown university: Nelson Francis & Henry Kucera

1 million words, written American English

- The **LOB** corpus: 1970, Geoffrey Leech & colleagues, Lancaster University

Corpus of written British English → finished by 1978: **L**ancaster-**O**slo-**B**ergen

Ideal to compare the two, i.e. British vs. American

Corpus linguistics: past and present

Later, the **BROWN** „family” was extended further:

- **FROWN** (Freiberg-Brown): written American English from 1991
- **FLOB** (Freiberg-Lancaster-Oslo-Bergen): written British English from 1991
- **BE06**: written British English from the early years of the 21st century

Corpus linguistics: past and present

COBUILD

John Sinclair, University of Birmingham: what corpora could reveal about *meaning*.

Early project: English Lexical Studies (1963); corpus analysis to find out the relationship between words and meaning → meaning is best seen as a property of words in combination → collocation.

→ Collins Birmingham University International Language Database (COBUILD) project, 1980. Creation of Bank of English corpus → result:

- Collins COBUILD English Language Dictionary (1987) - Collins COBUILD English Dictionary (1995)

- Collins COBUILD English Grammar (1990)

Corpus linguistics: past and present → Future?

Continuous development of tools and technologies in CL

Corpus techniques are finding their own way to other disciplines

The broader field of Digital Humanities is developing rapidly

Corpus linguistics: past and present → Future?

- University of Nottingham, Headtalk project → construction of a corpus of *video streams* of an academic supervision session, marked-up to indicate aspects of *body language*. → relationship between language and gestures
- Lancaster University: CL and Geographical Information Systems → extracting place-names from a corpus, searching for their semantic collocates, and creating maps to allow users to visualise how certain concepts are distributed geographically.

Corpus-based versus corpus-driven linguistics

Corpus-**based** studies typically use corpus data in order to explore a theory or hypothesis, aiming to validate it, refute it or refine it. The definition of corpus linguistics as a *method* underpins this approach.

Corpus-**driven** linguistics rejects the characterisation of corpus linguistics as a method and claims instead that the corpus *itself* should be the sole source of our hypotheses about language. It is thus claimed that the corpus itself embodies a theory of language (Tognini-Bonelli 2001: 84-5).

Tools

- [AntConc – Download](#)
- [English Corpora: most widely used online corpora. Billions of words of data: free online access](#)
- [MICASE 50 / Corpus Class Test](#)
- [XML Helsinki Corpus Browser](#)

Corpus use in reality 😊

TASK 1.

OE ME EModE ???

Fæder – Fader – Father

Corpus use in reality 😊

[Aethelwold: The Benedictine Rule](#)

[l. 258](#): he sceal gedon, þæt his reðnes ateowe *Fæderlice* and

[l. 591](#): heofonlican *Fæder*.

[l. 766](#): and eac swa þa haligan trahtas fram namcuþum *Fæderum* and

[l. 1048](#): we rædað, þæt þa ealdan *Fæderas* on anum dæge þæt

[l. 1245](#): he ealre geferrædenne swa swa *Fæder*.

Corpus use in reality 😊

[l. 258](#): he sceal gedon, þæt his reðnes ateowe *Fæderlice* and

He must act so that his mercy shows forth the Fatherly

[l. 591](#): heofonlican *Fæder*.

heavenly Father

[l. 766](#): and eac swa þa haligan trahtas fram namcūpum *Fæderum* and

and likewise the holy commentaries by the well-known Fathers (of the Church).

[l. 1048](#): we rædað, þæt þa ealdan *Fæderas* on anum dæge þæt

We read that the ancient Fathers in one day

[l. 1245](#): he ealre geferrædenne swa swa *Fæder*.

He guides the whole community as a Father does.

Anonymous: Aelred of Rielvaux's De Institutione Inclusarum (3)

l. 51: loke þu hiere þe veys of þy **fader**, which seiþ to þe þat al þy ioye

listen to the voice of your Father, who says to you that all your joy...

l. 201: wydue godes, and **faderles** and moderles children godes, and also

the goods of widows, and the goods of fatherless and motherless children, and also...

Anonymous: Aelred of Rielvaux's De Institutione Inclusarum (4)

l. 36: and there may she here his **fader** speke to hir and say, as the ...

and there she may hear her Father speak to her and say, as the ...

l. 158: gretly thyn holy purpos for truly, as Seynt Gregory seith, holy **faders**

greatly your holy purpose, for truly, as Saint Gregory says, the holy Fathers...

l. 1250: Ye neither know me, nor my **Father**

230: Take these things hence, make not my **fathers** house an house of merchandize.

l. 386: 35 The **Father** loueth the Sonne,

But the houre commeth, and

4 7 3 now is, when the true worshippers

4 7 4 shall worship the **Father** in spirit, and

4 7 5 in trueth: for the **Father** seeketh such

4 7 6 to worship him.

Corpus use in reality ☺

Possible questions:

- **Spelling form** (e.g. *fæder, fader, father*)
- **Syntactic environment** (just the structure of the clause: OV? SV? V2?)
- **Semantic nuances** (How is the word being used? Literal? Metaphorical? Kinship? Religious?)

**Thank You
for your attention!**