KEMENTERIAN AGAMA REPUBLIK INDONESIA
INSTITUT AGAMA ISLAM NEGERI PONOROGO
FAKULTAS TARBIYAH DAN ILMU KEGURUAN
Terakreditasi B sesuai SK BAN PT Nomor:2619/SK/BAN-PT/Ak-SURV/PT/XI/2016
Alamat : Jl. Pramuka No.156 Po.Box. 116 Ponorogo 63471 Tlp. (0352) 481277 Fax. (0352) 461893
Fax. (0352) 461893 Website: www.iainponorogo.ac.id E-mail: www.info@iainponorogo.ac.id

SURAT TUGAS

Nomor: B-5513 /In.32.2/PP.00.9/10/2020

Bersama ini kami menugaskan Saudara :

1. Nama : **Dr. DHINUK PUSPITA KIRANA, M.Pd**
2. Jabatan : **Dosen Tetap PNS**
3. NIP : **198303272011012007**
4. Instansi : **Jurusan Tadris Bahasa Inggris Fakultas Tarbiyah dan Ilmu Keguruan Institut Agama Islam Negeri Ponorogo**

Untuk menjadi **narasumber/pembicara** pada acara Webinar yang berjudul ***"Getting Acquainted with Corpus Linguistics and the Tools"*** di Program Studi Pendidikan Bahasa Inggris Fakultas Keguruan dan Ilmu Pendidikan Universitas Mulawarman Samarinda Kalimantan Timur pada :

Hari : Selasa
Tanggal : 03 Nopember 2020
Acara : Webinar di Prodi Pend. Bhs Inggris FKIP Universitas Mulawarman Samarinda Kalimantan Timur

Demikian surat tugas ini dibuat untuk dapat dipergunakan sebagaimana mestinya.

Ponorogo, 15 Oktober 2020
Dekan,

KEMENTERIAN AGAMA REPUBLIK INDONESIA IAIN PONOROGO Fakultas Tarbiyah dan Ilmu Keguruan

AHMADI

English Language Education Study Program
Faculty of Teaching and Teacher Education
Mulawarman University

November 3, 2020

GETTING ACQUAINTED WITH CORPUS LINGUISTICS AND THE TOOLS

Dr. Dhinuk Puspita Kirana, M. Pd.
(IAIN Ponorogo, Jawa Timur)

Dr. Bibit Suhatmady, M. Pd.
(Universitas Mulawarman, Kalimantan Timur)

INTRODUCTION TO CORPUS LINGUISTICS

Dr. Dhinuk Puspita Kirana, M.Pd

Sekolah Tinggi Agama Islam Negeri Ponorogo, Ponorogo

Outline

1. • Linguistics
2. • Corpus
3. • Corpus Linguistics

STAIN Ponorogo since 2011

IAIN Ponorogo – now

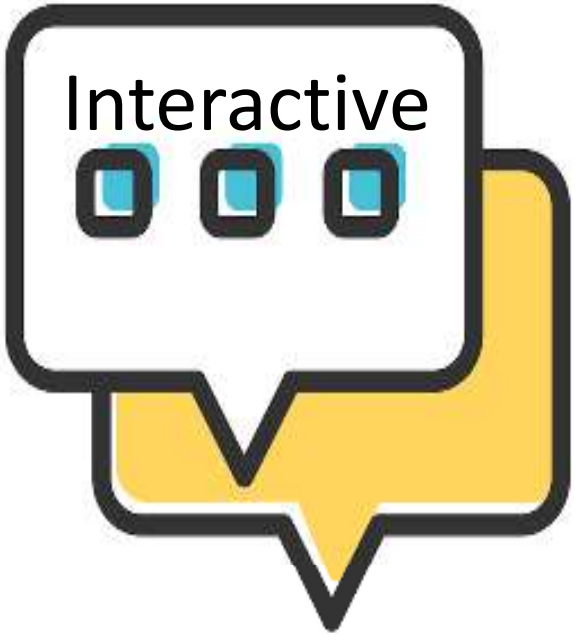
Research interests:

TEFL

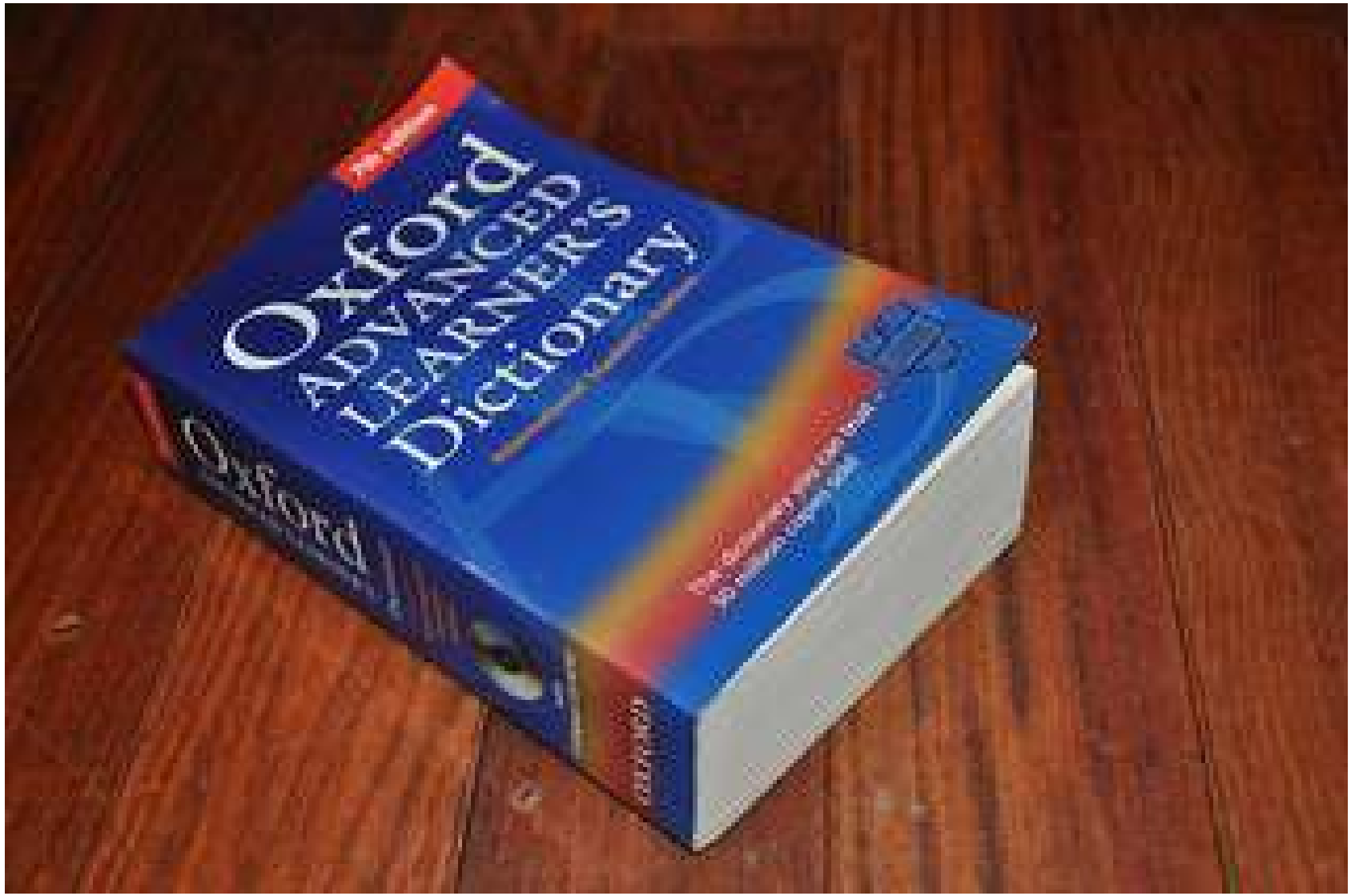
Vocabulary

Discourse Analysis

Corpus Linguistics

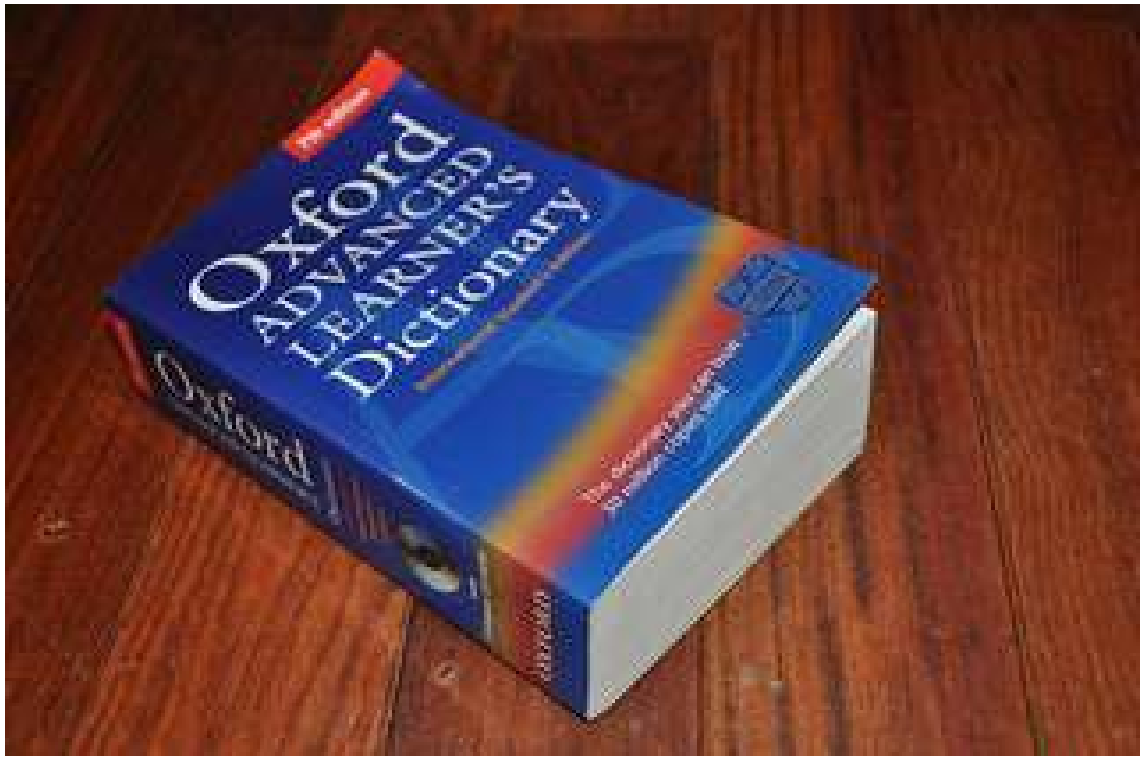


Communicate with me during the presentation through the chat box


Oxford
ADVANCED
LEARNER'S
Dictionary
Oxford

- How many words should you know from these sea of words?

Which of these words should you learn first before the others?



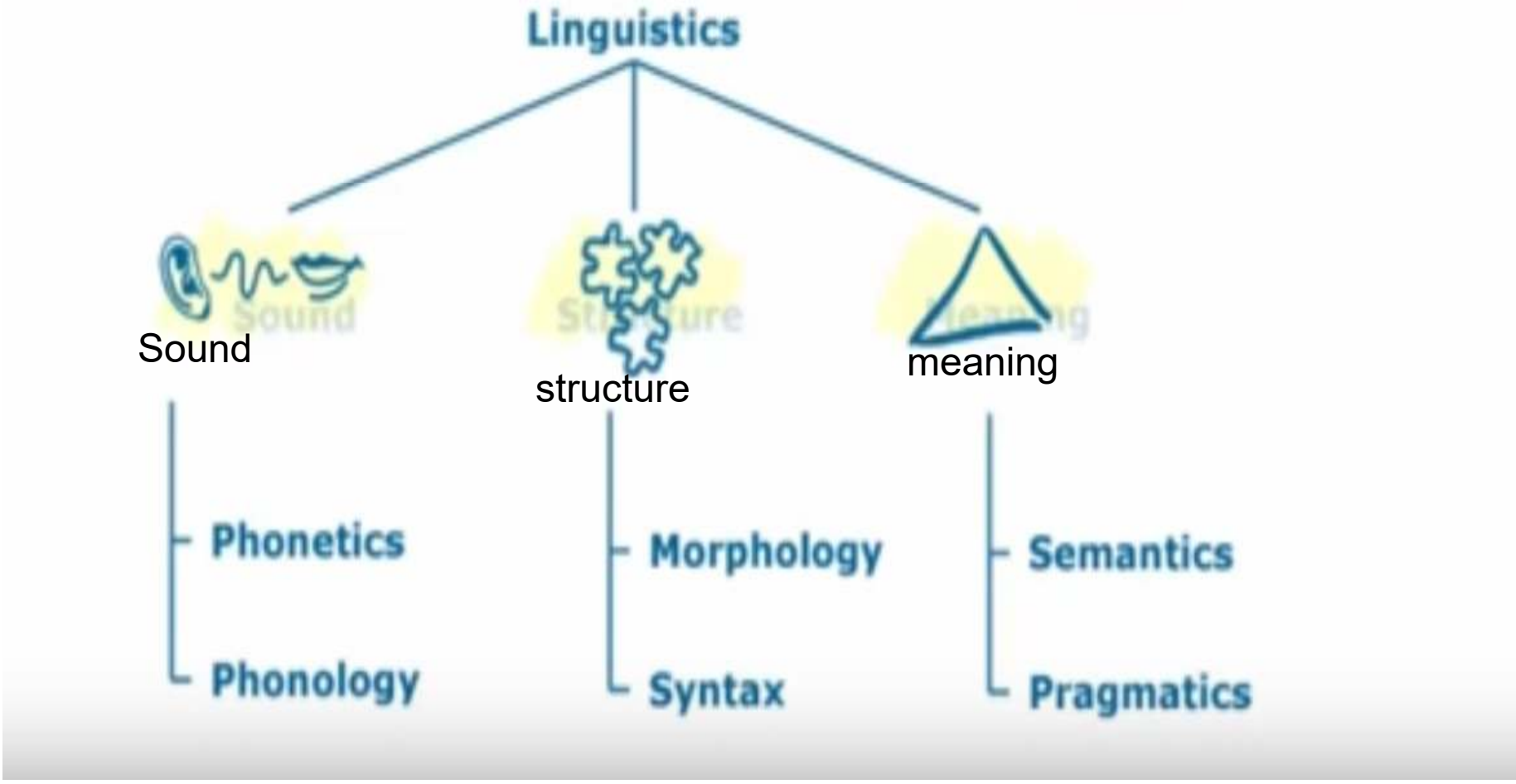
What is Linguistics

Linguistics

- the scientific study of language.

- **Language:** the principal method of human communication, consisting of words used in a structured and conventional way and conveyed by speech, writing, or gesture (Oxford Dictionary)


psycho-
computational-
socio-
forensic
clinical
neuro-
corpus-


Linguistics
Sound
structure
meaning
Phonetics
Phonology
Morphology
Syntax
Semantics
Pragmatics

What is Corpus

Comparing adjectives change over time

Whether two-syllable words had an “-er” or “more”

corpus

What is Corpus

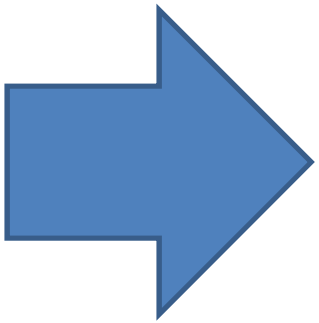
CORPUS (Latin word) : body

- "A collection of
- (1) *machine-readable*
- (2) *authentic* texts (including transcripts of spoken data)
- (3) *sampled* to be
- (4) *representative* of a particular language or language variety."

"**Corpora**"

plural form of "**corpus**"

- Corpus: large collections of texts which represent a sample of a particular variety or use of language(s) that are presented in machine readable form.



Corpus Linguistics

A way of investigating language by observing large amounts of naturally-occuring, electronically-stored discourse, using software which selects, sorts, matches, counts and calculates".

(Hunston, Susan and Gill Francis, 2000).

Why is corpus linguistics important?

- to see how language is used,
- To see how that language is used in different contexts,
- to learn/teach language more effectively.

Most Popular Words

	I	**to**
a	**is**	
of	**in**	
the		**and**
That	**it**	

- **Elt Irl**

Most Popular Words

	I	**to**
a	**is**	
of	**in**	
the		**and**
That	**it**	

- **Elt Irl**

Most Popular Words

	I	**to**
a	**is**	
of	**in**	
the		**and**
That	**it**	

- **Elt Irl**

Most Popular Words

I **to**

a **is**

of **in**

the **and**

That **it**

- **Elt Irl**

Most Popular Words

	I	**to**
a		**is**
of	**in**	
the		**and**
That	**it**	

- **Elt Irl**

How popular the word is?

- INDONESIA

BNC (100 million words)

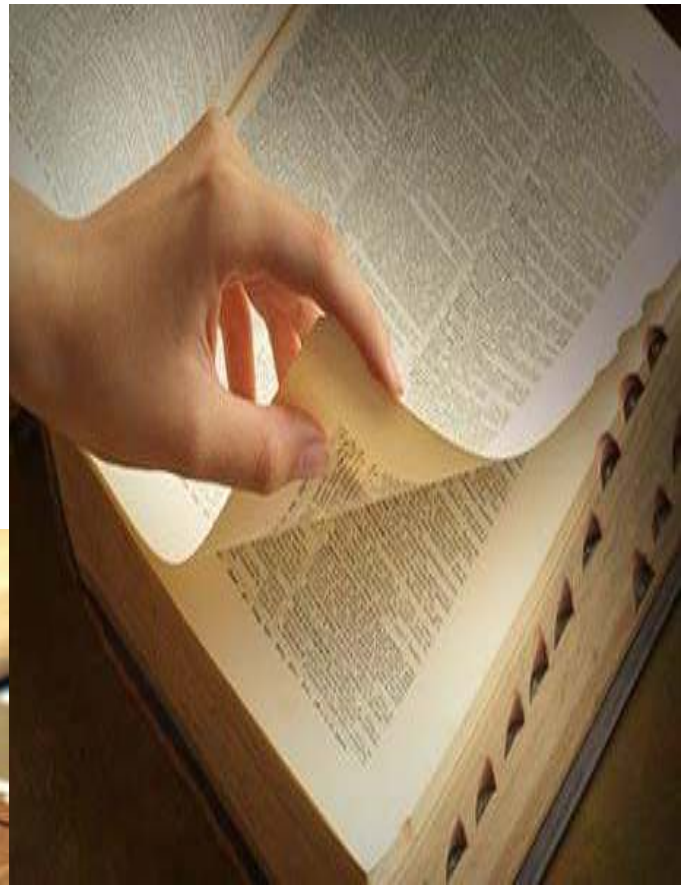
the	5.973.437
of	3.009.801
and	2.587.880
to	2.565.070
a	2.136.923

Indonesia	824

BNC

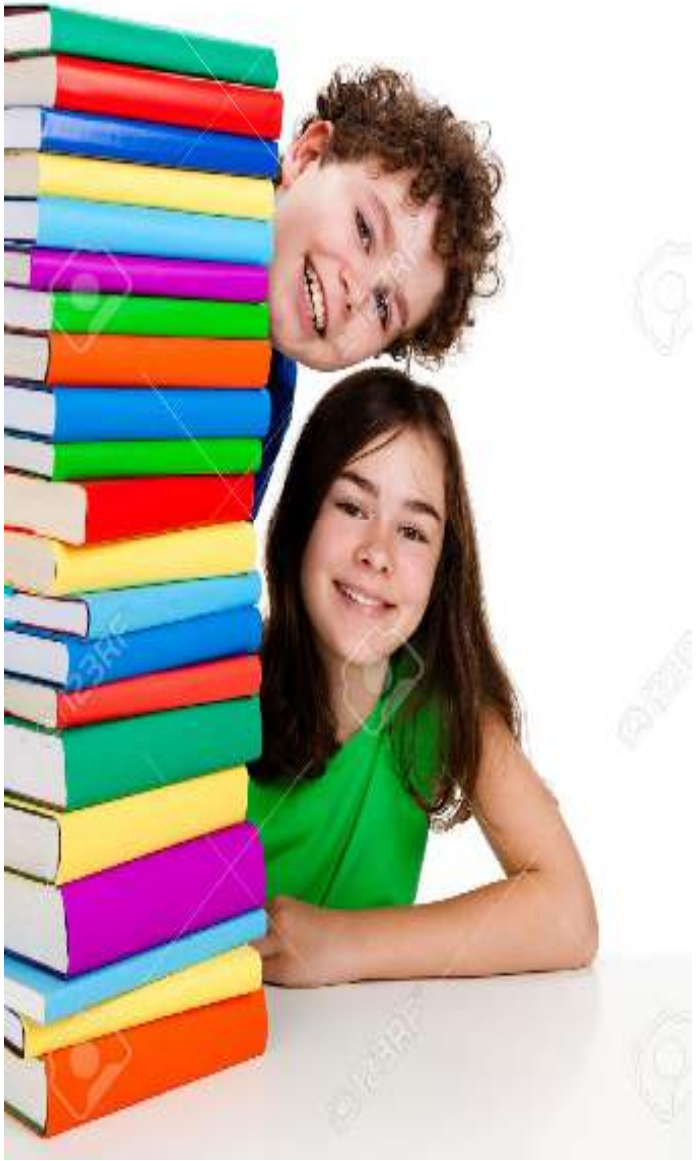
- The British National Corpus (BNC)
- Oxford University press
- in the 1980s - early 1990s,
- 100 million words of text texts from a wide range of genres
- (e.g. spoken, fiction, magazines, newspapers, and academic).

- Which words language learner should learn first before the others?



- Beginner language learners are suggested to set the initial goal at the high-frequent vocabulary of 2,000 words.

- The 2,000 most frequent words are fundamental for any language use (Nation, 1990).



- For university EFL learners, knowing the most frequently used 2,000 words will give them the competence to communicate effectively in speaking and writing modes.

- How the linguists know these word list?

Corpus-Based Approach

- Corpus-based approach is empirical, analyzing the actual patterns of use from natural texts. It utilizes a large and principled collection of natural texts as the basis for analysis.
- It makes extensive use of computers for analysis, using both automatic and interactive techniques.
- It integrates both quantitative and qualitative analytical techniques

Word Lists	
One	Represents the 1st 1000 most frequent words in English
Two	Represents the 2nd 1000 most frequent words in English
Three	Includes words that are not found in the first 2,000 words but are frequent in secondary school and university texts. These lists are based on Michael West's General Service List (1953) and on Coxhead's Academic Word List (1998).
Not in the list	The words which are not included in the three lists. This list may include less frequent than 14,000-word level, proper nouns, acronyms, abbreviations, alternative spelling, letters with numbers, exclamations, errors and non-English words.


thank you
danke
рахмат
謝謝
ngiyabonga
teşekkür ederim
gracias
Баярлалаа
спасибо
faafetai lava
mersi
kia ora
barka
welalin
tack
spas
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
vinaka
спасибі
blagodaram
kiitos
dankie
nanni
nandri
dhanyavad
mauruuru
köszönöm
bedankt
bayarlalaa
gràcie
hvala
enkosi
dziękuję
akun
dankon
aciu
хвала
asante
mana
obrigada
murakoze
djiere dieuf
tau
mochchakkeram
дякую
mamnun
chokrane
go raibh maith agat
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
sobodi
dėkuji
obrigado
sagolun
sukriya
kop khun krap
arigatô
takk
dakujem
trugarez
mèsi
didi madloba
ありがとう
najis tuke
tanemirt
grazie
kam sah hamnida
rahmat
terima kasih
rahmet
ευχαριστώ
diolch
dhanyavadagalu
shukriya
mercê
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
merci

Spoken Corpora

- Recording
- Transcribing, coding, mark-up
- Management and analysis

(Pstathas and Anderson, 1990; Leech et al., 1995; Lapadat and Lindsay, 1999; Thompson, 2005; Knight et al., 2006

Analysis Procedure

Nature of corpus-based approach

- It is empirical, analysing the actual patterns of use from natural texts
- It utilises a **large and principled collection** of natural texts as the basis for analysis
- It makes extensive use of **computers** for analysis, using both automatic and interactive techniques
- It integrates **both quantitative and qualitative** analytical techniques

(Biber et al 1998: 4-5)

What is a corpus?

- The word *corpus* comes from Latin (“body”) and the plural is *corpora*
- A corpus is a body of naturally occurring language
  - …but rarely a random collection of text
  - Corpora “are generally assembled with particular purposes in mind, and are often assembled to be (informally speaking) *representative* of some language or text type.”

- “A corpus is a collection of (1) *machine-readable* (2) *authentic* texts (including transcripts of spoken data) which is (3) *sampled* to be (4) *representative* of a particular language or language variety.”

What is a corpus for?

- A corpus is made for the study of language in a broad sense
  - To test existing linguistic theory and hypotheses
  - To generate and verify new linguistic hypotheses
  - Beyond linguistics, to provide textual evidence in text-based humanities and social sciences subjects
- The purpose is reflected in a well-designed corpus

Why use corpora?

- Even expert speakers have only a partial knowledge of a language
  - A corpus can be more comprehensive and balanced
- Even expert speakers tend to notice the unusual and think of what is possible
  - A corpus can show us what is common and typical
- Even expert speakers cannot quantify their knowledge of language
  - A corpus can readily give us accurate statistics

Why use corpora?

- Even expert speakers cannot remember everything they know
  - A corpus can store and recall all the information that has been stored in it
- Even experts speakers cannot make up natural examples
  - A corpus can provide us with a vast number of examples in real communication context
- Even expert speakers have prejudices and preferences and every language has cultural connotations and underlying ideology
  - A corpus can give you more objective evidence

Why use corpora?

- Even expert speakers are not always available to be consulted
  - A corpus can be made permanently accessible to all
- Even expert speakers cannot keep up with language change
  - A constantly updated corpus can reflect even recent changes in the language
- Even expert speakers lack authority: they can be challenged by other expert speakers
  - A corpus can encompass the actual language use of many expert speakers

References

- https://www.youtube.com/watch?v=bzz1pFWAtMo
- https://www.youtube.com/watch?v=32RjJ-lA-8Q