

ΓΕ77  
COMPUTATIONAL LINGUISTICS

**Athanasios N. Karasimos**

***akarasimos@gmail.com***

BA in Linguistics | National and Kapodistrian University of Athens

Lecture 9 | Wed 23 May 2018

Lecture 10 | Wed 30 May 2018

# CORPUS LINGUISTICS I

## Introduction to Corpus Linguistics

# OUTLINE

- Lectures 9 + 10: Introduction to Corpus linguistics
  - What is and is not a corpus?
  - Why use corpora?
  - Corpora vs. intuitions
  - The corpus methodology
  - A brief history of Corpus Linguistics
  - Nature and applications of corpus-based studies
- **TASK:** testing your intuitions + exploring online resources

# CORPUS: DEFINITION

# 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.” (Leech 1992)
- “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.” (MXT 2006: 5)

# WHAT IS NOT A CORPUS?

- A list of **words** is not a corpus
  - Building blocks of language
- A **text archive** is not a corpus
  - A random collection of texts
- A collection of **citations** is not a corpus
  - A short quotation which contains a word or phrase that is the reason for its selection
- A collection of **quotations** is not a corpus
  - A short selection from a text chosen on internal criteria by human beings
- A **text** is not a corpus
  - Intending to be read in different ways
- The **Web** is not a corpus
  - Its dimensions unknown, constantly changing, not designed from a linguistic perspective

Sinclair (2005)

# 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*

# THE SCOPE OF CORPUS LINGUISTICS

- Corpus makers or compilers.
- Developers of tools for the analysis of corpora.
- Descriptive linguists.
- Exploiters of corpus-based linguistic descriptions for use in a variety of applications such as language learning and teaching, natural language processing by machine, including speech recognition and translation.

# THE OTHER SIDE OF INTUITION

- Intuitions are always useful in linguistics
  - To invent (grammatical, ungrammatical, or questionable) example sentences for linguistic analysis
  - To make judgments about the acceptability / grammaticality or meaning of an expression
  - To help with categorization

# THE OTHER SIDE OF INTUITION

Intuitions should be applied with caution

- Possibly biased as they are likely to be influenced by one's dialect or sociolect
- Introspective data is artificial and may not represent typical language use as one is consciously monitoring one's language production
- Introspective data is decontextualized because it exists in the analyst's mind rather than in any real linguistic context
- Intuitions are not observable and verifiable by everyone as corpora are
- Excessive reliance on intuitions blinds the analyst to the realities of language usage because we tend to notice the unusual but overlook the commonplace
- There are areas in linguistics where intuitions cannot be used reliably e.g. language variation, historical linguistics, register and style, first and second language acquisition
- Human beings have only the vaguest notion of the frequency of a construct or a word

# BENEFITS OF CORPUS DATA

- Corpus data is more reliable
  - A corpus pools together linguistic intuitions of a range of language speakers, which offsets the potential biases in intuitions of individual speakers
- Corpus data is more natural
  - It is used in real communications instead of being invented specifically for linguistic analysis
- Corpus data is contextualized
  - Attested language use which has already occurred in real linguistic context
- Corpus data is quantitative
  - Corpora can provide frequencies and statistics readily
- Corpus data can find differences that intuitions alone cannot perceive
  - E.g. synonyms *totally, absolutely, utterly, completely, entirely*

# CORPORA VS. INTUITIONS

- Not necessarily antagonistic, but rather corroborate each other and can be gainfully viewed as being complementary
  - Armchair linguists and corpus linguists “need each other. Or better, [...] the two kinds of linguists, wherever possible, should exist in the same body.” (Fillmore 1992)
  - “Neither the corpus linguist of the 1950s, who rejected intuitions, nor the general linguist of the 1960s, who rejected corpus data, was able to achieve the interaction of data coverage and the insight that characterize the many successful corpus analyses of recent years.” (Leech 1991)
- The key to using corpus data is to find the balance between the use of corpus data and the use of one’s intuitions

# THE CORPUS METHODOLOGY

- It is debatable whether CorLing is a methodology or a branch of linguistics
  - CorLing goes well beyond this methodological role and has become an independent discipline
- In spite of the name, CorLing is indeed a methodology rather than an independent branch of linguistics in the same sense as phonetics, syntax, semantics or pragmatics
  - These latter areas of linguistics describe, or explain, a certain aspect of language use
  - Corpus linguistics, in contrast, is not restricted to a particular aspect of language - it can be employed to explore almost any area of linguistic research



# THE HISTORY OF CORPUS

A brief history of Corpus Linguistics

## A BRIEF HISTORY OF CORLING

- The term *corpus linguistics* first appeared only in the early 1980s, but corpus-based language study has a substantial history
- The history of CorLing can be split into two periods: before and after Chomsky

# B.C. HISTORY

- Before Chomsky
  - Field linguists and linguists of the structuralist tradition used “shoebox corpora” – shoeboxes filled with paper slips
    - Their methodology was essentially “corpus-based” in the sense that it was empirical and based on observed data
  - The work of early corpus linguistics was underpinned by two fundamental, yet flawed assumptions
    - The sentences of a natural language are finite.
    - The sentences of a natural language can be collected and enumerated.
  - Most linguists saw the “corpus” as the only source of linguistic evidence in the formation of linguistic theories

# CHOMSKY'S REVOLUTION

- **Chomsky revolution:** Between 1957 and 1965 Chomsky changed the direction of linguistics from empiricism towards rationalism
  - “Any natural corpus will be skewed. Some sentences won’t occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list.” (Chomsky 1962)
  - Our internal knowledge of language in human brain (competence) replaces observed data (performance)
  - Intuitions started to be relied on as evidence

Xiao, R. (2008)

# THE RETURN OF CORLING

- Revival of CorLing
  - Corpus research was continued in a few centres (Brown, Lancaster) in the 60s-70s
    - The Brown University Standard Corpus of Present-day American English (Brown corpus)
    - Lancaster-Oslo-Bergen Corpus of BrE (LOB)
  - The hardware still imposed some restrictions until the real development started in the 1980s
    - The relation of corpora with computer technology rekindled interest in the corpus methodology
    - Since then, the number and size of corpora and corpus-based studies have increased dramatically
  - Nowadays, the corpus methodology enjoys widespread popularity, and has opened up or foregrounded many new areas of research

# CORPUS LINGUISTICS AND THE OTHERS

The Interaction of Corpus Linguistics

# AREAS THAT HAVE USED CORPORA

- Lexicography
- Lexical studies
- Grammatical studies
- Register/genre analysis
- Language variation
- Contrastive analysis
- Translation studies
- Language change
- Language teaching
- Semantics
- Pragmatics
- Stylistics
- Literary study
- Sociolinguistics
- Discourse analysis
- Forensic linguistics
- Computational linguistics
- ...

# 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)



# THE COMPUTER POWER

- Development of computer technology has revived CL
- Machine-readability is a de facto attribute of modern corpora
- Electronic corpora have advantages unavailable to their “shoebox” ancestors
  - It is the use of computerized corpora, together with computer programs which facilitate linguistic analysis, that distinguishes modern electronic corpora from early ‘drawer-cum-slip’ corpora

# THE COMPUTER POWER

- Computerized corpora can be processed and manipulated rapidly at minimal cost
  - E.g. searching, selecting, sorting and formatting
- Computers can process machine-readable data accurately and consistently
- Computers can avoid human bias in an analysis, thus making the result more reliable
- Machine-readability allows further automatic processing to be performed on the corpus so that corpus texts can be enriched with various metadata and linguistic analyses
  - Corpus markup and corpus annotation

# QUESTIONING DEEP THOUGHT

- “Alright,” said the computer Deep Thought. “The Answer to the Great Question...”
- “Yes...!”
- “Of Life, the Universe and Everything ...” said Deep Thought.
- “Yes...!”
- “Is...”
- “Yes...!!!...?”
- “Forty-two,” said Deep Thought, with infinite majesty and calm.
- It was a long time before anyone spoke.
- “Forty-two!” yelled someone in the audience. “Is that all you’ve got to show for seven and a half million years’ work?”
- “I checked it very thoroughly,” said the computer, “and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you’ve never actually known what the question is.”

Hitchhikers Guide to the Galaxy by Douglas Adams

# WHAT CORPORA CANNOT DO

- Corpora do not provide negative evidence
  - Cannot tell us what is possible or not possible
  - Can show what is central and typical in language
- Corpora can yield findings but rarely provide explanations for what is observed
  - Interfacing other methodologies
- The use of corpora as a methodology also defines the boundaries of any given study
  - Importance of amenable research questions
- The findings based on a particular corpus only tell us what is true in that corpus
  - Generalization vs. representativeness

# ASKING THE RIGHT QUESTION

- Corpus linguistics as a methodology is only one of the (many) ways of doing things – “doing linguistics”
- The usefulness of corpora depends upon the research question being investigated
  - “They are invaluable for doing what they do, and what they do not do must be done in another way.” (Hunston 2002: 20)
- The development of the corpus-based approach as a tool in language studies has been compared to the invention of telescopes in astronomy
  - If it is ridiculous to criticize a telescope for not being a microscope, it is equally pointless to criticize the corpus-based approach for not doing what it is not intended to do
- It is up to you to formulate research questions amenable to corpus-based investigation and to decide how to combine corpora with other resources

# BRITISH NATIONAL CORPUS

**Corpus Linguistics Lab I:**  
*Testing your intuitions with BUY-BNC*

# MOST COMMON NOUN IN ENGLISH

- Search for [n\*]

List  Chart  Collocates  Compare  KWIC

[POS]

Sections  Texts/Virtual  Sort/Limit  Options

	<input type="checkbox"/>	CONTEXT	ALL <input type="checkbox"/>	SPOKEN <input type="checkbox"/>	FICTION <input type="checkbox"/>	MAGAZINE <input type="checkbox"/>	NEWSPAPER <input type="checkbox"/>	NON-ACAD <input type="checkbox"/>	ACADEMIC <input type="checkbox"/>	MISC <input type="checkbox"/>
1	<input type="checkbox"/>	TIME	142575	17959	29301	10280	14074	21049	17516	32396
2	<input type="checkbox"/>	PEOPLE	117821	20753	13123	7293	14949	21560	14431	25712
3	<input type="checkbox"/>	WAY	89704	12308	20812	6421	7450	12303	12575	17835
4	<input type="checkbox"/>	YEARS	85992	6774	8496	8024	14198	15864	10660	21976
5	<input type="checkbox"/>	YEAR	69604	7455	3293	7645	16240	11020	4770	19181
6	<input type="checkbox"/>	WORK	58836	4766	5200	4540	4562	11930	11815	16023
7	<input type="checkbox"/>	GOVERNMENT	58193	2319	690	2716	8353	18811	11161	14143
8	<input type="checkbox"/>	DAY	55428	7016	12650	4660	6923	7329	3298	13552
9	<input type="checkbox"/>	MAN	55314	3677	23028	3040	7272	5306	4440	8551
10	<input type="checkbox"/>	WORLD	54279	1524	6240	6251	8605	11041	7329	13289
11	<input type="checkbox"/>	LIFE	52083	2527	10742	3812	5494	9329	7395	12784

# MOST COMMON NOUN IN ADVERTS

- Search [nn\*] in Section W\_advert

List Chart Collocates Compare KWIC

[nn\*]

Sections  Texts/Virtual  Sort/Limit  Options

1 W\_religion  
W\_admin  
**W\_advert**  
W\_biography  
W\_commerce  
W\_email  
W\_essay\_schl

2 IGNORE  
-----  
SPOKEN  
FICTION  
MAGAZINE  
NEWSPAPER  
NON-ACAD

	<input type="checkbox"/>	CONTEXT	FREQ	
1	<input type="checkbox"/>	HOTEL	1009	
2	<input type="checkbox"/>	WORLD	788	
3	<input type="checkbox"/>	TIME	784	
4	<input type="checkbox"/>	CENTRE	765	
5	<input type="checkbox"/>	HOLIDAY	760	
6	<input type="checkbox"/>	DAY	734	
7	<input type="checkbox"/>	SERVICE	665	
8	<input type="checkbox"/>	YEAR	601	
9	<input type="checkbox"/>	RANGE	584	
10	<input type="checkbox"/>	CLUB	576	
11	<input type="checkbox"/>	YEARS	554	
12	<input type="checkbox"/>	FACILITIES	552	
13	<input type="checkbox"/>	HOUSE	542	
14	<input type="checkbox"/>	BAR	540	
15	<input type="checkbox"/>	INFORMATION	505	



# ADJECTIVES: FICTION VS. NON-FICTION

List Chart Collocates Compare KWIC

Sections Texts/Virtual Sort/Limit Options

1

2

SORTING  **SEC1 : SEC2**

MINIMUM   5  0

SEC 1 (MISC): 20,835,159 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	AGGREGATE	873	0	41.9	0.0	4,190.0
2	REGULATORY	527	0	25.3	0.0	2,529.4
3	OFFLINE	442	0	21.2	0.0	2,121.4
4	KEYNESIAN	323	0	15.5	0.0	1,550.3
5	NON-EXECUTIVE	171	0	8.2	0.0	820.7
6	TAXABLE	171	0	8.2	0.0	820.7
7	MACROECONOMIC	158	0	7.6	0.0	758.3
8	NO-ARBITRAGE	148	0	7.1	0.0	710.3
9	NATIONALISED	139	0	6.7	0.0	667.1
10	SHORT-RUN	133	0	6.4	0.0	638.3

SEC 2 (FICTION): 15,909,312 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SABINE	458	17	28.8	0.8	35.3
2	HUSKY	170	7	10.7	0.3	31.8
3	FLUSHED	102	5	6.4	0.2	26.7
4	CLAMMY	78	4	4.9	0.2	25.5
5	GREY-HAIRED	55	3	3.5	0.1	24.0
6	RUEFUL	116	7	7.3	0.3	21.7
7	MUTTERED	75	5	4.7	0.2	19.6
8	BLASTED	41	3	2.6	0.1	17.9
9	COLD-BLOODED	53	4	3.3	0.2	17.4
10	WORDLESS	52	4	3.3	0.2	17.0

# TALK[V] VS. TALK[N]

SECTION (CLICK FOR SUB-SECTIONS) (SEE ALL SECTIONS AT ONCE)	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
<a href="#">SPOKEN</a>	399	10.0	40.05	
<a href="#">FICTION</a>	1,093	15.9	68.70	
<a href="#">MAGAZINE</a>	368	7.3	50.67	
<a href="#">NEWSPAPER</a>	496	10.5	47.39	
<a href="#">NON-ACAD</a>	513	16.5	31.10	
<a href="#">ACADEMIC</a>	439	15.3	28.63	
<a href="#">MISC</a>	812	20.8	38.97	

SECTION (CLICK FOR SUB-SECTIONS) (SEE ALL SECTIONS AT ONCE)	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
<a href="#">SPOKEN</a>	2,742	10.0	275.20	
<a href="#">FICTION</a>	4,212	15.9	264.75	
<a href="#">MAGAZINE</a>	638	7.3	87.85	
<a href="#">NEWSPAPER</a>	899	10.5	85.89	
<a href="#">NON-ACAD</a>	1,113	16.5	67.47	
<a href="#">ACADEMIC</a>	716	15.3	46.70	
<a href="#">MISC</a>	1,579	20.8	75.79	

1	D97	S_meeting	A B C	. (SP:D97PS002) (unclear) when was that? (SP:D97PSUNK) (unclear) (SP:D97PS003) Wasn't that on the <b>talk</b> ? Do you remember (unclear) this talk we had (unclear) th
2	D97	S_meeting	A B C	(SP:D97PSUNK) (unclear) (SP:D97PS003) Wasn't that on the talk? Do you remember (unclear) this <b>talk</b> we had (unclear) that was last year. Er (SP:D97PS002) (unclea
3	D97	S_meeting	A B C	got (pause) what they've got on here is they've got (pause) whale watch <b>talk</b> and slides. (unclear). Punch and Judy show. Magic (unclear) and juggling with
4	DCH	S_meeting	A B C	time who'd been to Central America recently and she gave us a very interesting <b>talk</b> on a visit to El Salvador and Guatamala, erm, and we, we
5	DCH	S_meeting	A B C	's in hand. (SP:DCHPSUNK) Oh, I see she hasn't actually given a <b>talk</b> , but she's going to. But she's, she going to deal
6	DCH	S_meeting	A B C	the way erm, the erm, the other few points were erm Jackie's <b>talk</b> last month she mentioned that erm she was gon na give sort of the more

1	D97	S_meeting	A B C	. (SP:D97PS002) (unclear) when was that? (SP:D97PSUNK) (unclear) (SP:D97PS003) Wasn't that on the <b>talk</b> ? Do you remember (unclear) this talk we had (unclear) th
2	D97	S_meeting	A B C	(SP:D97PSUNK) (unclear) (SP:D97PS003) Wasn't that on the talk? Do you remember (unclear) this <b>talk</b> we had (unclear) that was last year. Er (SP:D97PS002) (unclea
3	D97	S_meeting	A B C	got (pause) what they've got on here is they've got (pause) whale watch <b>talk</b> and slides. (unclear). Punch and Judy show. Magic (unclear) and juggling with
4	DCH	S_meeting	A B C	time who'd been to Central America recently and she gave us a very interesting <b>talk</b> on a visit to El Salvador and Guatamala, erm, and we, we
5	DCH	S_meeting	A B C	's in hand. (SP:DCHPSUNK) Oh, I see she hasn't actually given a <b>talk</b> , but she's going to. But she's, she going to deal
6	DCH	S_meeting	A B C	the way erm, the erm, the other few points were erm Jackie's <b>talk</b> last month she mentioned that erm she was gon na give sort of the more

# THE BATTLE OF SYNONYMS: SMART VS. CLEVER

List Chart Collocates **Compare** KWIC

smart Word1 [POS]  
clever Word2 [POS]  
\* Collocates [POS]

+ 4 3 2 1 0 0 1 2 3 4 +

Compare words Reset

Sections Texts/Virtual Sort/Limit Opti

WORD 1 (W1): SMART (0.78)						WORD 2 (W2): CLEVER (1.29)					
	WORD	W1	W2	W1/W2	SCORE		WORD	W2	W1	W2/W1	SCORE
1	SUIT	32	0	64.0	82.4	1	BOY	67	1	67.0	52.0
2	DRIVE	19	0	38.0	48.9	2	HALF	24	1	24.0	18.6
3	CLOTHES	19	0	38.0	48.9	3	BIT	12	0	24.0	18.6
4	BLACK	18	0	36.0	46.4	4	DICK	11	0	22.0	17.1
5	CARD	18	0	36.0	46.4	5	TRICKS	11	0	22.0	17.1
6	DRESS	16	0	32.0	41.2	6	FOLLY	10	0	20.0	15.5
7	SUITS	16	0	32.0	41.2	7	HANDS	10	0	20.0	15.5
8	LOOKING	30	1	30.0	38.6	8	HOWEVER	10	0	20.0	15.5
9	SHOES	15	0	30.0	38.6	9	CLEVER	38	2	19.0	14.8
10	UNIFORM	13	0	26.0	33.5	10	MARKETING	9	0	18.0	14.0

# WORD COLLOCATION

List Chart **Collocates** Compare KWIC

caused Word/phrase [POS]

\_nn\* Collocates noun.ALL

+ 4 3 2 1 0 0 1 2 3 4 +

Find collocates Reset

Sections Texts/Virtual Sort/Limit Options

SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]

	<input type="checkbox"/>	CONTEXT	FREQ	
1	<input type="checkbox"/>	DAMAGE	641	
2	<input type="checkbox"/>	PROBLEMS	360	
3	<input type="checkbox"/>	LOSS	148	
4	<input type="checkbox"/>	DEATH	134	
5	<input type="checkbox"/>	INJURY	104	
6	<input type="checkbox"/>	TROUBLE	104	
7	<input type="checkbox"/>	FIRE	100	
8	<input type="checkbox"/>	PROBLEM	90	
9	<input type="checkbox"/>	CONCERN	88	
10	<input type="checkbox"/>	ACCIDENT	81	

COMPARE

# DATA: SINGULAR OR PLURAL?

List Chart **Collocates** Compare KWIC

data Word/phrase [POS]

\_v?z\* Collocates verb.3SG

+ 4 3 2 1 0 0 1 2 3 4 +

Find collocates Reset

List Chart **Collocates** Compare KWIC

data Word/phrase [POS]

\_vv0\* Collocates verb.BASE

+ 4 3 2 1 0 0 1 2 3 4 +

Find collocates Reset

we can prove it today. (SP:PS21B) Mm mm. (SP:PS21C) Which **is** based on historical **data** but (SP:PS21B) Mm mm. (SP:PS21C) d  
there we are. (SP:PS4W3) Yeah the other the other factor **is** now that that **data is** getting on for twelve months old (SP:PS4W2)  
own monitoring that you won't get much variation. It **provides** us with the **data**, but school children will probably lose interest  
What they've done in a matter of a few weeks **is** put together a **data** bank of information that **is** colossal, and they're dealing w  
by switching labour-intensive software engineering work to India. West Germany **needs** about 4,000 electronic **data** processin  
around to your heart **is** content. Of course, it **doesn't** recover any **data**, but it **does** wonders for your temper. The UK supplier,  
important reading on US inflation later today when the Labour Department **is** due to release **data** on its US Consumer Price In  
' interview † respondents formally (van Maanen 1982: 140 **argues** that most ethnographic **data** are conversation-based). As on  
record **is** transferred to his next school. If the record **is** computerised, the **Data** Protection Act 1984 **does** not give any right to  
strong tendency for manual workers to live close to their workplace **is** further demonstrated by **data** on some of the larger loc  
reported in Swann (p. 60) but little interpretive use **is** made of these **data**. What must always have been clear to perceptive te  
to determining which † kind † of pupil † succeeds and which **does** not. This **data is** incorporated in publicizing of the unit and its  
the use of the same key-presses for the same effects and **allows** the transfer of **data** between functions.) There **is** insufficient  
or other of the two main parties. The third generalization **is** drawn from survey **data**. Gallup Poll data showed that in these fo  
large amount of space; the ceramic of an FRAM chip **is** capable of storing **data** in a much smaller space. Ramtron aims in futu  
of the sky, the Infrared Astronomical Satellite (IRAS) **is** turning in high-quality **data** in such profusion that astronomers are hav  
running time †. What **is** really needed to clinch things **is** not just more **data**, but some sign of the Zo particle, which, if the elect  
, which **uses** length of whale; yet the Japanese model **does** not fit the **data** as well as the IIED model. They also refuse to accep  
's Rutherford-Appleton Laboratory and Carlo Rubbia from CERN. The team **has** been scrutinising new **data** collected during A  
that a French company **plans** to launch in 1984. IBM **has** processed simulated SPOT **data** (from high-flying aircraft) to analyse

you how we can compute Chow tests in Microfit if you **come** out of the **data** processing environment type Q erm and **move** to the action menu I guess  
and **generate** a dummy variable (pause) right, so if you **go** into the erm **data** processing environment (pause) if it's in the er sort of **process** plot (unde  
I believe Bob (----) has asked you to er, I **think**, **collect** some **data** erm, on trade in wheat and cotton erm, as an example. We  
a problem? (SP:HE7PSUNK) Oh yes, yes, obviously we **need** to **keep** our **data** secure. (SP:PS2U4) Well it's a threat to your personal safety. Absolutely,  
, I'll save the **file**, and er, I **want** to copy that **data**, let's **say** to here. **Put** the cell point in there. If  
you know, and they would actually **work** over (pause) er **work** on erm (pause) **data** erm, from their computer (pause) and order er erm (pause) you kn  
won't be of any benefit unti-, until you actually **start** working on true **data**. But in the meantime we will **work** on company averages for you, until  
animal experiments if that's the case? Surely they only **take** place when the **data's** known to be applicable to people.' She sighed heavily, as if  
there's your table. If, later on, you **want** to **change** the **data** in the table, you can do it in your document without having to **go**  
just this. If you type: TREE /F > PENGUIN ENTER then all the **data** from the TREE command, which normally goes straight to the screen, is redirected  
coating has a higher saturation current value, requiring a higher **write** current. So **data** written at the lower DD rate will not saturate the HD magnetic  
COPY device to **read** each **file** in turn. If you **want** to check the **data** on and entire disk **complete** with sub-directories then try XCOPY \*, \* NUL /S  
CAD part files. These sub-structures **give** efficient and fast information **recall** by partitioning the **data base** into a defined structure. The design data ba  
NDT and CM techniques remain inefficient. This is because they **generate** excessive amounts of **data** and information that must be interpreted by spe  
necessary for masking, etc. In this section we briefly **discuss** the forms of **data transfer** instruction commonly found on word-oriented computers: first  
They perform the following procedures: # (i) # **input** the digitized map **data** to GIMMS, and # (ii) # **generate** the linkages between line segments  
databases is given in this section. Cartographic data manipulation operations **include** # transformation of **data** from one map projection or scale to an  
, the application of geography to real-world problems. These applications **require** the assembling of **data** and concepts from the different systematic b

# NOUN DESCRIPTION

List Chart Collocates **Compare** KWIC

Word1 [POS]: woman

Word2 [POS]: man

Collocates: j\* adj.ALL

+ 4 3 2 1 0 0 1 2 3 4 +

WAIT... Reset

Sections Texts/Virtual **Sort/Limit** Options

SORTING: RELEVANCE WORD1: WORD2

MINIMUM: FREQUENCY  5 0

WORD 1 (W1): WOMAN (0.37)						WORD 2 (W2): MAN (2.67)					
	WORD	W1	W2	W1/W2	SCORE		WORD	W2	W1	W2/W1	SCORE
1	DUMPY	12	0	24.0	64.1	1	UTD	72	0	144.0	53.9
2	LIBERATED	10	0	20.0	53.4	2	BEARDED	70	0	140.0	52.4
3	MENSTRUATING	10	0	20.0	53.4	3	UNITED	33	0	66.0	24.7
4	MOTHERLY	10	0	20.0	53.4	4	MACHO	32	0	64.0	24.0
5	WEEPING	9	0	18.0	48.0	5	BURLY	30	0	60.0	22.5
6	SAMARITAN	8	0	16.0	42.7	6	SELF-MADE	29	0	58.0	21.7
7	TOKEN	8	0	16.0	42.7	7	RIGHT-HAND	55	1	55.0	20.6
8	PRETTY	59	4	14.8	39.4	8	UNDERGROUND	27	0	54.0	20.2
9	HYSTERICAL	14	1	14.0	37.4	9	ARMED	22	0	44.0	16.5
10	HINDU	7	0	14.0	37.4	10	BALDING	21	0	42.0	15.7

# TESTING PHRASAL VERBS: REASON FOR/ TO

List Chart **Collocates** Compare KWIC

reason for Word/phrase [POS]

Collocates

+ 4 3 2 1 0 0 1 2 3 4 +

why I put it there. (SP:PS23F) I accept everything. (SP:PS23B) There was some **reason for having** it there. It's cos we knew that otherwise you'd be back and see how it had changed. (SP:PS273) Mm. Mm. (SP:PS26Y) As a **reason for going**. (SP:PS273) Yeah. I mean I think that in terms of visits (SP:HYKPSUNK) Right, move approval. (SP:HYKPSUNK) Yes. (unclear) (SP:HYKPSUNK) Has anyone else **giving reason for** these (unclear) (SP:PS3CH) I . (unclear) I mean you must see it yourself. Okay there must be a **reason for doing** this, but er I'm more concerned with **getting oil out** (unclear) live near their daughter who was in (----) and we felt that was a legitimate **reason for paying** a higher rate and, and, and we did do so, not beyond it and I'm not proposing to alter that. The whole br-- **reason for bringing** the scheme forward in the programme was associated with th him (SP:J45PSUNK) Yes, I do represent him. Erm, Mr (----) explained the **reason for changing** this but basically it's to, to swap us again and that key points. (SP:J97PSUNK) 'Cos otherwise everyone (pause) (SP:J97PSUNK) Everyone will (unclear)2. (SP:J97PSUNK) The **reason for setting** aside a (u closely to that, and only vary them when they've got a perfectly good **reason for doing** so. In paragraph nine, I report that erm, members of some A CT capacity within one of the subsidiaries within the group um the **reason for highlighting** it **highlighting** it is not particularly to make a sor can speak out, damn the consequences. (SP:PS5M9) I think public relations is the **reason for doing** it isn't it and to pacify the politicians, (SP:K6WPS the state institutions really. (SP:PS6GB) It doesn't (SP:PS6GC) (unclear) (SP:PS6GB) Part of the **reason for doing** it is is we can actually say we did it, s what transpired. So, I think we're no nearer forward to **having a reason for** the explosion, but erm there hasn't been one since. (SP:PS6H9) Any danger is when **having** a good time is the reason for **living** and the only **reason for** it, you see, if god has intervened in our life, if , which had not included the scissors' business with Sister Mary as **being** the **reason for her leaving** because that might have put any caring mothe

List Chart **Collocates** Compare KWIC

reason to Word/phrase [POS]

Collocates

+ 4 3 2 1 0 0 1 2 3 4 +

're pissed off it doesn't, you **don't have to have** a special **reason to be** pissed off you can just **be** pissed off **like** you can just **be** do it for money, I mean I **think** probably (unclear) to **be** a better **reason to become** a councillor these days. It affects your job, your job prospects in all things. Some of us here today may not **have** all that much **reason to rejoice** right now but God will ultimately lead us all forward in joy. information that (----) was in the master bedroom is true, they might **have** every **reason to come** into the master bedroom to **see** if er (pause) the, who was rates consistent with the approved strategy. And we **don't see** in principle any **reason to** divert from them. The second point I **want to make** is in relation and the garden disappears. For ever. She had reason to **be** wary, **reason to act** cool. She'd almost lost her garden of paradise in heat and clear emotion was that he should not **know**. She would never **allow** him a **reason to** pity her again, to hold her in contempt. It was one of and political implications of their actions. Government intervention would inevitably **follow** if Governments had **reason to believe** that their interests were Ariadne had any intention of going anywhere but it's nice to **have** a solid **reason to stay put**. Talbot thought briefly. 'Solves one little problem, to Randall Lodge together, and to her chagrin she could **think** of no good **reason to refuse**. With luck, he would need all his concentration for the traffic to speculate, was still surprised by her arrival. He could **think** of no **reason to account** for it. Over the past few years, since she had broken thought you might **want to see** it." There must **have** been a **reason to come** here for a day." They met with the president. him about the snake. Had she done so, he might **have** had a **reason to object** to the holiday and a great deal of trouble would **have** been saved this he grasped on to it with relief for it seemed to **give** him a **reason to do** nothing, though in his heart he knew it was **fear**, not couldn't **have** been a sough or drain, for there could **have** been no **reason to** drain water into a mine. Joseph Usher, Tace's hero, had

# CORPUS TASKS

- 1) What are the top 5 modal verbs in English?
- 2) Is there any difference between verbs *destroy*, *ruin*, and *demolish*? If so, what is it?
- 3) Do you think the adjectives in “utterly + adjective” have anything in common? If so what is that?
- 4) Can we use the plural form of *research* as in “his researches”?



# CORPUS LINGUISTICS II

A second approach of Corpora

# BEST KNOWN CORPORA

- The Birmingham Collection of English Texts (COBUILD)
- The Bank of English
- The British National Corpus (BNC)
- The Brown Corpus
- The Lancaster-Oslo/Bergen Corpus (LOB)
- The Helsinki Corpus of English Texts: Diachronic and Dialectal
- The International Corpus of English (ICE)
- The Old English of New Zealand (ONZE)
- Scottish Corpus of Texts and Speech (SCOTS)

# CHOMSKY VS. CORPUS LINGUISTICS

- Chomsky criticizes Corpus Linguistics
  - Frequency tells you about the world rather than about language (the sentence I live in New York is fundamentally more likely than I live in Dayton Ohio).
  - Corpus research is slow and limited.
  - Corpus leaves out what you don't say, which can be more informative than what you say.
  - Pseudo-techniques.

## RE: CHOMSKY VS. CORPUS LINGUISTICS

- Performance is still an inherently valid object of study. Entire fields of science and research use exclusively or almost exclusively observational data: astronomy, archeology, paleontology, biology, etc.
- Naturally-occurring data can be collected, studied, analysed, commented and referred to. Corpus-based observations are more verifiable than introspectively based statements.
- The finite-infinite is not a big issue, since in many other fields we also have an infinite number of possible examples, but it does not stop us from studying them.
- A big enough corpus (such as a 100 million word British National Corpus) will provide a lot of utterances one is likely to encounter in language.

## RE: CHOMSKY VS. CORPUS LINGUISTICS

- Frequency lists compiled objectively from corpora have shown that human intuition about language is very specific and far from being a reliable source.
- Word frequency is also a good reason to use very large and well-balanced corpora.
- Corpora are now collected in extremely systematic and controlled ways.
- Corpus analysis will never tell you that an utterance is impossible. But with a large enough and well balanced corpus and sufficient statistical tools, it can tell you when it is statistically significant for such an utterance to be absent from the corpus.

# THE “PROPERTIES” OF CORPORA

- Authenticity
- Objectivity
- Verifiability
- Exposure to large amounts of data
- New insights into language
- Enhancement of learner motivation

# CORPORA: AUTHENTICITY

- Key notion in the field of corpus work.
- “One does not study all of botany by artificial flowers” (Sinclair 1991:24).

# CORPORA: OBJECTIVITY

- No prior selection of data.
- “I am above all an observer; I quite simply cannot help making linguistic observations. In conversations at home and abroad, in railway compartments, when passing people in streets and on roads, I am constantly noticing oddities of pronunciation, forms and sentence constructions”. (Jespersen 1995: 213)



## CORPORA: VERIFIABILITY

- “Verifiability is a normal requirement in scientific research, therefore, the science of language – linguistics -- (which is often claimed to be the scientific study of language) should not be exempt from this standard mode of research procedure” (Leech 1991:112).

# CORPORA: LANGUAGE INSIGHTS

- Sinclair noted (1991:1) that “traditionally linguistics has been limited to what a single individual could experience and remember... Starved of adequate data, linguistics languished – indeed it became totally introverted. It became fashionable to look inwards to the mind rather than outwards to society. Intuition was the key, and similarity of language structure to various formal models was emphasized. The communicative role of language was hardly referred to.... Students of linguistics over many years have been urged to rely heavily on their intuition and to prefer their intuitions to actual text where there is some discrepancy. Their study has, therefore, been more about intuition than about language”.
- Many subtle observations.
- Corpora can help learners discover new meanings of the words they already know.
- New understanding of meaning in Corpus Linguistics.

# CORPORA: MOTIVATION

- “Corpus as an information source fits in very well with the dominant trend in university teaching philosophy over the past 20 years, which is the trend from teaching as imparting knowledge to teaching as mediated learning” (Leech 1997:2).
- There is no longer a gulf between research and teaching, since the student is placed in a position similar to that of a researcher, investigating and imaginatively making sense of the data available through observation of the corpus.
- McCarthy (1998: 67-68) argues that the traditional ‘Three Ps’ methodology Presentation – Practice – Production should be supplemented by the ‘Three Is’ method: Illustration – Interaction – Induction.
- Students “discover” language.

# CORPORA: MOTIVATION

- The potential value in foreign language teaching is considerable for at least 2 reasons:
- The first is the Hawthorne effect – a well-known principle according to which any **new** tool or method tends to stimulate the actors of a pedagogic act and to improve the results more than the mere continuance of trite procedures.
- The second is connected with the Laws of memory: memory is conditioned by an active cognition of the past.
- Recognizing and recalling a word are in the long run much easier if the mind, at the very moment of the input, has actively associated the fragment with circumstances of that input.

# HUGE AMOUNTS OF DATA

- Nurtures a “feel of language”, develops an understanding of what is natural in a language.
- The computer is “ a tireless native-speaker informant, with rather greater potential knowledge of the language than the average native speaker” (Barnbrook 1996: 140).

## DISADVANTAGES OF USING CORPORA

- A corpus is not an infallible source of all linguistic information about language.
- Overdependence and overreliance upon corpora can be an inhibiting dogma.
- An attempt to replace a laborious hands-on analysis by a rapid automatic processing.

# CORPUS CREATION

# CORPUS CREATION I

- The issues in corpus design and compilation are directly related to the validity and reliability of the research based on a particular corpus (Kennedy 1998: 60).
- Sinclair (1991: 13) claimed that “the decisions that are taken about what is to be in the corpus, and how the selection is to be organized, control almost everything that happens subsequently. The results are only as good as the corpus”.



## CORPUS CREATION II

- Getting permissions
- Discussion and research points.
- Research the copyright laws of Greece and find out what restrictions govern the production of an electronic copy of copyrighted material for research purposes. Contact one or more publishers to find out about their policy and practice in assisting researchers to build corpora.
- Further reading (McEnery *et al.* 2006: 77-79)

## CORPUS CREATION III

- The design of a corpus is dependent upon the type of a corpus and purpose for which the corpus is to be used.
- Types of corpora (sample, monitor, general, spoken, written, learner, translation, parallel, comparable, etc).

# SAMPLE CORPORA

- A sample corpus is a static collection of texts (samples of texts) selected according to some strict criteria and intended to be typical of the whole language or an aspect of the language at a particular period of time.
- Brown and LOB corpora consist of a large number (500) short extracts (2000 words), randomly selected from within 15 genres of printed texts.

# MONITOR CORPORA

- Monitor corpora are text corpora that represent a dynamic, changing picture of a language. Such a dynamic collection of texts is constantly growing and changing with the addition of new text samples.

# GENERAL CORPORA

- They are assembled to serve as a reference base for unspecified linguistic research (Kennedy 1998:19).
- The **size** of a corpus: as a general rule, the bigger a corpus is the richer and more interesting the output from a concordancing program will be, and the more likely to represent accurately features of the language.

# SPOKEN AND WRITTEN CORPORA

- The spoken form of the language is a better guide to the **fundamental organization** of the language than the written form.
- Spoken language is primary and all the changes start there.
- Spoken language is not that well researched.
- Spoken language can also prove valuable for the studies of differences between speech and writing.

# LEARNER CORPORA

- Learner corpora are defined as electronic collections of authentic texts produced by foreign or second language learners (Granger 2003).
- The first computerised learner corpora were collected in the 1990s when several learner corpora projects were launched: the Longman Learners' Corpus, the Cambridge Learner Corpus, the Hong Kong University Learner Corpus and the International Corpus of Learner English (ICLE).

# LEARNER CORPORA

- The [Longman Learners' Corpus](#) contains ten million words of text written by learners of English of different levels of proficiency and from twenty different L1 backgrounds.
- The [Cambridge Learner Corpus](#) is a large collection of written texts from learners of English all over the world.
- The [International Corpus of Learner English](#) (ICLE) is the best-known learner corpus which provides a collection of essays written by advanced learners of English (third and fourth year university students) from different native language backgrounds. ([v2.](#))



# LEARNER CORPORA

- Language acquisition is a mental process, which we can observe only through its product, i.e. the data the learner produces.
- Learner corpora can provide a wider empirical basis on which many hypotheses can be tested and the principles that govern the process of learning a foreign language uncovered.
- The introduction of corpora in the classroom might mean a tough job of changing attitudes of teachers and learners.
- Educating teachers and spreading the word about corpora.
- Using corpora in the classroom changes the student's role.
- “The distinction between teaching and research becomes blurred and irrelevant” (Knowles 1990).

# CORPORA IN TRANSLATION STUDIES

- The use of corpora in translation studies is relatively new - it was first advocated by Mona Baker in 1993.
- Linguists viewed translations with suspicion, assumed them to be ontologically different from non-translated texts and referred to them as 'interlanguage' (Selinker 1972), 'third language' (Duff 1981), 'third code' (Frawley 1984), or 'translationese' (e.g. Gellerstam 1986, Doherty 1998, Mauranen 1999, Tirkkonen-Condit 2002).

# PARALLEL CORPORA

- A parallel corpus is a corpus composed of source texts and their translations in one or more different languages; parallel corpora can be aligned at a word, phrase or sentence level thus establishing correspondences between units of bilingual or multilingual texts.
- Parallel corpora are important resources for translation studies. As Aijmer and Altenberg (1996:12) noted, they can provide new insights into the languages compared, insights that cannot be obtained in studies of mono-lingual corpora, they can also be used for different comparative purposes and enhance our understanding of language-specific, typological and cultural differences as well as universal features, they can highlight differences between source texts and translations, they can also be used for a number of practical applications in translation teaching.

# PARALLEL CORPORA

- A parallel corpus is a corpus composed of source texts and their translations in one or more different languages; parallel corpora can be aligned at a word, phrase or sentence level thus establishing correspondences between units of bilingual or multilingual texts.
- Parallel corpora are important resources for translation studies. As Aijmer and Altenberg (1996:12) noted, they can provide new insights into the languages compared, insights that cannot be obtained in studies of mono-lingual corpora, they can also be used for different comparative purposes and enhance our understanding of language-specific, typological and cultural differences as well as universal features, they can highlight differences between source texts and translations, they can also be used for a number of practical applications in translation teaching.

# COMPARABLE CORPORA

- Comparable corpora are comparable original texts in two or more languages, they are monolingual corpora designed using the same sampling techniques, e.g. the Aarhus corpus of contract law (McEnery 2006: 47).
- Monolingual comparable corpus is particularly useful in studying intrinsic features of translations, improving the translator's understanding of the subject domain, terminology and idiomatic expressions in the specific field.

# TRANSLATIONAL CORPORA

- Corpora may be integrated into translator training and may meet various needs of translator trainers.
- Parallel corpora are especially useful as they can be used to retrieve terminology, explore collocations, phrasal patterns, lexical polysemy, translation of collocations and idioms, etc. (Botley et al. 2000).
- The students can also be encouraged to compile their own specific corpora that can be very useful for content information, terminology, phraseology in some specific domains or topics.
- A corpus compilation experiment can be carried out as a real-life translation assignment.

# TRANSLATIONAL CORPORA

- Comparable corpora can also be helpful in translator training as they can be used to check terminology and collocates, identify text-type-specific formulations, validate intuitions and provide explanations for appropriateness of certain solutions to problems (Pearson 2003).
- Corpora can be very useful in translator's profession: specialized corpora can be used to familiarize translators with concepts and terms from a specific domain, translators can study corpora output to understand text-type conventions, literary translators can also resort to corpora data to study an author's style, to find some literary devices, etc.

# UNDERSTANDING OF MEANING

(Re)Viewing Meaning through



# SINCLAIR'S UNDERSTANDING OF MEANING

- The methodological steps proposed by Sinclair to identify what he calls “extended unit of meaning are:
- identify **collocational profile** (lexical realizations)
- identify **colligational patterns** (lexico-grammatical realizations)
- consider common semantic field (**semantic preference**)
- consider pragmatic realisations (**semantic prosody**)

## EXTENDED UNIT OF MEANING

- **Collocation** is the occurrence of words with no more than four intervening words.
- **Colligation** is the co-occurrence of grammatical phenomena, and on the syntagmatic axis our descriptive techniques at present confine us to the co-occurrence of a member of a grammatical class – say a word class- with a word or phrase.
- **Semantic preference** is the restriction of regular co-occurrence to items which share a semantic feature, for example that they are all about say, sport or suffering. Semantic preference is a semantic field a word's collocates predominantly belong to.

## EXTENDED UNIT OF MEANING

- **Semantic prosody** is attitudinal, and on the pragmatic side of the semantics/pragmatics continuum. Semantic prosody describes the way in which certain seemingly neutral words can be perceived with positive or negative associations through frequent occurrences with particular collocations. Thus, such verbs as **set in** (rot, decay, ill-will, decadence, infection, prejudice, etc.), **cause** (cancer, crisis, accident, delay, death, damage, trouble, etc.), **commit** (crime, offences, foul etc.), **rife** (crime, diseases, misery, corruption, speculation, etc.), often have negative semantic prosody, while such words as **impressive** will occur with lexical items such as **dignity, talent, gains, achievement**, etc. will have positive prosody.

# COLLOCATIONS

- First used by Firth (1957).
- “Collocations of a given word are statements of the habitual or customary places of that word” (Firth 1968: 181).
- Quantitative approach to collocations.
- “Collocations are not absolute or deterministic, but are probabilistic events, resulting from repeated combinations used and encountered by the speakers of any language” (O’Keefe et al. 2007: 59).
- Sinclair (1991) argues that there are two fundamental principles at work in the creation of meaning: the ‘idiom principle’ and the ‘open choice principle’.

# COLLOCATIONS

- Biber et al. (1991) refer to lexical bundles as recurrent strings of words, delimited by establishing frequency cut-off points, for example, that a string must occur at least 10 times per million words of text and must be distributed over a number of different texts.
- **Research points:**
- Use **BNCWeb** to analyse the collocations of the words of your choice.
- **Further reading:**
- McEnery *et al.* 2006

# IDIOMATICITY

- Different terminology: ‘lexical phrases’ (Nattinger and DeCarrico 1992), ‘prefabricated patterns’ (Hakuta 1974), ‘routine formulae’ (Coulmas 1979), ‘formulaic sequences’ (Wray 2002; Schmitt 2004), ‘lexicalized stems’ (Pawley and Syder 1983), ‘chunks’ (De Cock 2000) as well as the more conventionally understood labels such as ‘(restricted) collocations’, ‘fixed expressions’, multi-word units/ expressions’, ‘idioms’ etc.
- “Strings of more than one word whose syntactic, lexical and phonological form is to a greater or lesser degree fixed and whose semantics and pragmatic functions are opaque and specialised, also to a greater or lesser degree” (O’Keefe 2007: 80).
- ‘Idiom-prone’ words: body parts, money, light, colour and other basic notions.

# IDIOMATICITY

- ‘Paradox’ of idiomaticity: the very thing which for native speakers promotes ease of processing and fluent production seems to present non-native users with an insurmountable obstacle.
- Idioms are difficult to get right.
- Idioms can sound strange on the lips of non-native users.
- Idioms do not just ‘pop up’ in native speech; rather they occur as part of a more extended phenomenon that generates subtle webs of semantic, pragmatic and discourse prosodies.

# REFERENCES I

- Aijmer, K. 2009. *Corpora and Language Teaching*. Amsterdam/Philadelphia. J. Benjamins.
- Aijmer, K., B. Altenberg. 1991. *English Corpus Linguistics*. Longman: London and New York.
- Altenberg, B. 1998. 'On the phraseology of spoken English: the evidence of recurrent word combinations' in Cowie, A.P. (ed) *Phraseology: Theory Analysis and Applications*. Oxford: Oxford University Press.
- Altenberg, B. and Granger, S. 2001. Grammatical and lexical patterning of *make* in student writing', *Applied Linguistics* 22(2): 173-194.
- Aprijaskytė, R. and E. Pareigytė, 1982. *Some lexical Difficulties for the Lithuanian Learner of English*. Vilnius.
- Aston, G. And Burnard, L. 1998. *The BNC Handbook*. Edinburgh: Edinburgh University Press.
- Baker, M. 1993. 'Corpus linguistics and translation studies: Implications and Applications' in M. Baker, G. Francis and E. Tognini-Bonelli (eds) *Text and technology: In Honour of John Sinclair*. Amsterdam: John Benjamins, pp. 17-45.
- Baker, M. 1996. 'Corpus-based translation studies: The challenges that lie ahead'. In H. Somers (ed.) *Terminology, LSP and Translation: Studies in language engineering, in honour of Juan C. Sager*. Amsterdam, John Benjamins, pp. 175-86.
- Barnbrook, G. 1996. *Language and Computers: A Practical introduction to the Computer Analysis of language*". Edinburgh University Press.
- Biber, D. 1990. 'Methodological issues regarding corpus-based analyses of linguistic variation'. *Literary and Linguistic Computing* 5: 257-269.
- Blum-Kulka, S. 1986. 'Shifts of cohesion and coherence in translation', in J. House and S. Blum-Kulka (eds) *Interlingual and Intercultural Communication: Discourse and cognition in translation and second language acquisition studies*. Tübingen: Gunter Narr, pp. 17-35.



## REFERENCES II

- Botley, S.P., A.M. McEnery and A. Wilson. 2000. *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi.
- Butler, C. 1992. *Computers and Written Texts*. Oxford: Blackwell.
- Chomsky, N. 1957. *Syntactic Structure*. The Hague: Mouton.
- Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, Mass: MIT Press.
- Crowdy, S. 1993. Spoken Corpus Design. *Literary and Linguistic Computing*. Vol. 8, no. 4, p. 259-265.
- De Cock, S., Granger, S., Leech, G. and McEnery, T. 1998. 'An automated approach to the phrasicon of EFL learners' in Granger, S. (ed) *Learner English on Computer*. London: Longman, 67-79.
- De Cock, S. 2000. 'Repetitive phrasal chunkiness and advanced EFL speech and writing' in Mair, C. and Hundt, M. (eds) *Corpus Linguistics and Linguistic Theory*. Papers from ICAME 20 1999. Amsterdam: Rodopi, 51-68.
- Coulmas, F. 1979. 'On the sociolinguistic relevance of routine formulae', *Journal of Pragmatics* 3: 239-66.
- Dagneaux E., Denness S., Granger S. And Meunier, F. 1996. *Error Tagging Manual Version 1.1*. Centre for English Corpus Linguistics. Université Catholique de Louvain, Louvain-la-Neuve.
- Doherty, M. 1998. 'Clauses or phrases – a principled account of *when*- clauses in translations between English and German', in S. Johansson and S. Oksefjell (eds) *Corpora and Cross-linguistic Research*, Amsterdam: Rodopi, pp. 235-54.
- Duff, A. 1981. *The Third Language: Recurrent problems of translation into English*, Oxford: Pergamon Press.
- Fillmore, Ch. 1992. "Corpus linguistics" or "Computer-aided armchair linguistics". In *Directions in Corpus Linguistics*. Proceedings of Nobel Symposium 82. 35-60.

## REFERENCES III

- Firth, J. 1957. *Papers in Linguistics*. Oxford: Oxford university Press.
- Frawley, W. 1984. 'Prolegomenon to a theory of translation', in W. Frawley (ed.) *Translation: Literary, linguistic and philosophical perspectives*, Newark, MD: University of Delaware Press, pp. 159-75; reprinted in L. Venuti (ed.) 2000. *The Translation Studies Reader*, London: Routledge, pp. 250-63.
- Granger, S. 2002. 'A bird's-eye view of learner corpus research' in S. Granger, J. Hung and S.Petch-Tyson (eds) *Computer learner Corpora, Second Language Acquisition and Foreign Language Teaching*, pp.3-33, Philadelphia: John Benjamins.
- Gass S.M. and Selinker, L. 2001. *Second Language Acquisition. An Introductory Course*. Mahwah NJ: Lawrence Erlbaum.
- Gellerstam, M. 1986. 'Translationese in Swedish novels translated from English', in L. Wollin and Landquist (eds) *Translation Studies in Scandinavia*, Lund: Gleerup, pp. 88-95.
- Granger S. 1996. 'From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora'. In K. Aijmer, B. Altenberg and M. Johansson (eds) *Language in Contrast: Papers from a Symposium on Text-Based Cross-Linguistic Studies*. Lund: Lund University Press.
- Granger, S. (ed). 1998. *Learner English on Computer*. London: Longman.
- Granger, S. 2003. International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly* 37 (3), 538-546.

## REFERENCES IV

- Granger, S. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching. A critical evaluation. In K. Aijmer (ed.) *Corpora and Language Teaching*. Amsterdam/Philadelphia: J. Benjamins. 13 – 32.
- Hakuta, K. 1974. 'Prefabricated patterns and the emergence of structure in second language acquisition', *Language Learning* 24: 287-298.
- Hermans, T. 1999. *Translation in Systems: Descriptive and system-oriented approaches explained*. Manchester: St Jerome
- Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Jespersen, O. 1995. *A Linguist's Life: An English Translation of Otto Jespersen's Autobiography*. ed. by A. Juul, et al. Odense: University Press of Southern Denmark.
- Johansson, S. 1998. 'On the role of corpora in cross-linguistic research'. In S. Johansson and S. Oksefjell (eds) *Corpora and Cross-linguistic Research*. Amsterdam: Rodopi, pp. 3-24.
- Johansson, S. 2007. *Seeing through Multilingual Corpora. On the use of Corpora in Contrastive Studies*. Amsterdam/Philadelphia: Benjamins.
- Johansson, S. and H. Hasselgård. 1999. Corpora and cross-linguistic research in the Nordic countries. In Granger et al (eds). *Contrastive Linguistics and Translation*, 145-162.

## REFERENCES V

- Johansson, S and K. Hofland. 2000. 'The English-Norwegian parallel corpus: current work and new directions', in S.P.Botley, A.M.McEnery and A.Wilson (eds) *Multilingual Corpora in Teaching and Research*, Amsterdam: Rodopi, pp. 134-47.
- Johns, T. 1991. ' "Should you be persuaded": two samples of data-driven learning materials' in T. Johns and P. King (eds) *Classroom concordancing ELR Journal 4*. University of Birmingham.
- Kaszubski, P. 1998., ,Enhancing a writing textbook: a national perspective' in S. Granger (ed) *Learner English on Computer*, pp. 172-185.
- Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London and New York: Longman.
- Kenny, D. 2001. *Lexis and Creativity in Translation: A corpus-based study*. Manchester: St Jerome.
- Knowles, G. 1990. The use of spoken and written corpora in the teaching of language and linguistics. *Literary and Linguistic Computing*.
- Laviosa, S. 2002. *Corpus-based Translation Studies: Theory, findings, applications*. Amsterdam: Rodopi.
- Leech, G. 1992. 100 million words of English: the British National Corpus (BNC). *Language Research*. Vol. 28, 1- 13.

## REFERENCES VI

- Leech, G. 1997. 'Teaching and language corpora: a convergence' in *Teaching and language Corpora*. ed. by A. Wichmann et al. Longman.
- Leech, G. 1998. 'Learner corpora: what they are and what can be done with them'. In *Learner English on Computer*. London: Longman, xiv-xx.
- Leech, G. And Fligelstone, S. 1992. 'Computers and corpus analysis' in Butler (ed.) *Computers and Written Texts*, pp. 115-140. Oxford: Blackwell.
- Mackin, R. 1978. On collocations: 'words shall be known by the company they keep'. P. Stevens (ed.) *Studies in Honour of A.S.Hornby*. Oxford: Oxford University Press. 149-164.
- McCarthy, M. 1998. *Spoken English and Applied Linguistics*. Cambridge: Cambridge University Press.
- McEnery, A. and T.Wilson (eds.) 1997. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- McEnery T., Xiao R. And Yukio Tono. 2006. *Corpus-Based Language Studies. An Advanced Resource Book*. London and New York: Routledge.
- Meyer, C. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

## REFERENCES VII

- Mukherjee, J and Rohrbach, J. 2006. Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research. In *Planning, Gluing and painting Corpora: Inside the Applied Corpus Linguist's Workshop*, B. Kettemann and G. Marko (eds.), 205-232. Frankfurt; Lang.
- Mauranen, A. 1999. 'Will "translationese" ruin a contrastive study?', *Languages in Contrast* 2:161-85.
- Nattinger, J. and DeCarrico, J. 1992. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.
- Nesselhauf, N. 2004. 'Learner corpora and their potential for language teaching'. In *How to Use Corpora in Language Teaching*. Ed. By J.M. Sinclair. Amsterdam/Philadelphia: J. Benjamins.
- O'Keefe, A. and M. McCarthy (eds). 2010. *The Routledge Handbook of Corpus Linguistics*. London: Routledge.
- O'Keefe, A., McCarthy M. and Carter R. 2007. *From Corpus to Classroom: language use and language teaching*. Cambridge: Cambridge University Press.
- Olohan, M. 2006. *Introducing Corpora in Translation Studies*. London and New York: Routledge.
- Olohan, M. and M. Baker 2000. 'Reporting *that* in translated English: Evidence for subconscious processes of explicitation?' *Across Languages and Cultures* 1: 141-72.
- Overas, L. 1998. 'In search of the third code: An investigation of norms in literary translation'. *Meta* 43: 571-88.
- Pawley, A. and Syder, F. 1983. 'Two puzzles for linguistic theory: nativelike selection and nativelike fluency' in Richards, J. and Schmidt, R. (eds) *Language and Communication*. New York: Longman, 191-226.
- Pearson, J. 2003. 'Using parallel texts in the translator training environment'. In F. Zanettin, S. Bernardini and D. Stewart (eds) *Corpora in Translator Education*. Manchester: St Jerome, pp. 15-24.

## REFERENCES VIII

- Pinker, S. 1994. *The Language Instinct*. New York: HarperCollins.
- Schmitt, N. 2004. *Formulaic Sequences*. Amsterdam: John Benjamins.
- Sinclair, J. 2000. Lexical Grammar. *Darbai ir Dienos*, t. 24, 191-203.
- Sinclair, J. 1991. *Corpus Concordance Collocation*. Oxford: Oxford University Press.
- Sinclair, J. (ed) 1996. *Looking Up. An Account of the COBUILD Project*. HarperCollins Publishers.
- Stewart, D. 2000. 'Poor relations and black sheep in translation studies', *Target* 12: 205-28.
- Stubbs, M. 2001. 'Texts, corpora, and problems of interpretation: a response to Widdowson'. *Applied Linguistics*. 22/2:149-172.
- Tirkkonen-Condit, S. 2002. 'Translationese: a myth or an empirical fact? A study into the linguistic identifiability of translated language', *Target* 14: 207-20.
- Toury, G. 1995. *Descriptive Translation Studies and Beyond*, Amsterdam: John Benjamins.
- Tognini-Bonelli, E. 2000. Corpus Classroom Currency. *Darbai ir Dienos* t. 24, 205-243.
- Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam/Philadelphia: J. Benjamins

## REFERENCES IX

- Vanderauwera, R. 1985. *Dutch Novels Translated into English: The transformation of a 'minority' literature*. Amsterdam: Rodopi.
- Varantola, K. 2003. 'Translators and disposable corpora'. In F. Zanettin, S. Bernardini and D. Stewart (eds) *Corpora in Translator Education*. Manchester: St Jerome, pp. 55-70.
- Widdowson, H. 2000. 'The limitations of linguistics applied'. *Applied Linguistics* 21/1:3-25.
- Williams, M. 1988. 'Language taught for meetings and language used in meetings: Is there anything in common?' *Applied Linguistics* 9 (1): 45-58.
- Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.
- Xiao, R. and M. Yue. 2012. 'Using corpora in Translation Studies: The state of the art'. In P. Baker. *Contemporary Corpus Linguistics*. New York: Continuum. 237-261.
- Zanettin, F., S. Bernardini and D. Stewart (eds). 2003. *Corpora in Translator Education*. Manchester: St. Jerome.
- Zipf, G.K. 1935. *The Psychobiology of Language*. Boston: Houghton Mifflin.
- LDEL, 1992. *Longman Dictionary of English Language and Culture*. Longman



# READINGS

- [Xiao, R. (2008) “Theory-driven corpus research: using corpora to inform aspect theory”. In A. Lüdeling & M. Kytö (eds.) *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter]