Corpus linguistics: a general introduction

What is Corpus Linguistics?

Corpus Linguistics is the study of language/linguistic phenomena through the analysis of data obtained from a corpus.

Theoretical aspects

Corpus linguistics

> "can be seen as a ***pre-application methodology***. […] by "pre-application" we mean that, unlike other applications that start by accepting facts as *given*, **corpus linguistics is in a position to define its own sets of rules and pieces of knowledge *before* they are applied.** […] Corpus linguistics has, therefore, a theoretical status and because of this it is in a position to contribute specifically to other applications. (Tognini-Bonelli, *Corpus linguistics at work*, 2001:1)

Historical background

Phase 1 – before 1950s

Franz Boas and the American Structuralism. He compiles small *corpora* to analyse the phonological aspects of the Inuit language, adopting an empirical approach

Phase 2 – after 1950s

USA – Leonard Bloomfield's *verificationism*: rejects the mental approach to language in favour of an empirical one. Language studies must rely on the observation of facts.

UK – the Firthian tradition: J.R. Firth – M.A.K. Halliday – J. Sinclair

They draw back on Malinowski's *context of culture* and *context of situation*. Language is a real phenomenon, which makes sense only if it is considered in its real use, i.e. as *performance* rather than as *competence*.

Historical Background

Reaction to Chomsky's transformational- generative grammar (mid-20th)

Dualism between *competence* and *performance*

Distinction between *deep structures* (*competence*) and *surface structures* (*performance*)

Language has to focus on *competence* rather than on *performance*

In short, the chomskyan linguistics

- Rejects *corpus linguistics* since a *corpus* is a collection of external data (*performance*)
- Is based on *introspection* and *rationalism* vs. *empiricism*.

Historical and Theoretical Issues

Firth/Halliday/Sinclair reject any dualism and opt for a *monist* view of language.

Focus on *performance*

To sum up some aspects in *CL*:

- *Empiricism* and direct observation of real data
- *Performance*
- *Form* and *content* are indivisible -> *lexico-grammar approach* to language
- *Parole* is context- and time-related. *Langue* is abstract and a-temporal
- Use of computers to study *corpora* qualitatively and quantitatively.

What is a corpus

- In linguistics, **corpus** (plural *corpora*) is a large and structured set of texts (now usually electronically stored and processed). A corpus may contain single texts in single language (*monolingual corpus*) or text data in multiple languages (*multilingual corpus*). Multilingual corpora that have been specially formatted for side-by-side comparison are called *aligned parallel corpora*. (*Webster's Online Dictionary*)

- A corpus is a collection of naturally-occurring language text, chosen to characterize a state or variety of a language. (Sinclair, *Corpus, Concordance, Collocation*, 1991:171)

What is a corpus

- A corpus can be defined as a collection of texts assumed to be representative of a given language put together so that it can be used for linguistic analysis. Usually the assumption is that the language stored in a corpus is naturally-occurring, that it is gathered according to explicit design criteria, with a specific purpose in mind, and with a claim to represent larger chunks of language selected according to a specific typology. […] in general there is consensus that a corpus deals with natural, authentic language. (Tognini-Bonelli, *Corpus linguistics at work*, 2001:2)

What is a corpus

- A corpus is a collection of texts, designed for some purpose, usually teaching or research. […] A corpus is not something that a speaker does or knows, but something constructed by a researcher. It is a record of performance, usually of many different users, and designed to be studied, so that we can make inferences about typical language use. Because it provides methods of observing patterns of a type which have long been sensed by literary critics, but which have not been identified empirically, the computer-assisted study of large corpora can perhaps suggest a way out of the paradoxes of dualism. (Stubbs, *Words and Phrases*, 2002:239-40)

What is a *corpus*?

- [A corpus is] a subset of an ETL (Electronic Text Library) built according to explicit design criteria for a specific purpose (Atkins, Clear and Osler, "Corpus Design Criteria", in *Literary and Linguistic Computing*, 7.1, 1992:1-16)

- a corpus is taken to be a computerised collection of authentic texts, amenable to automatic or semiautomatic processing or analysis. The texts are selected according to explicit criteria in order to capture the regularities of a language, a language variety or a sub-language. (Tognini-Bonelli, *op. cit.*:55)

It follows that

Texts must be collected according to specific criteria: content/genre/typology/register, etc.;

Texts must be available in machine-readable form

Texts are collected in order to analyse specific linguistic phenomena

Criteria

Authenticity

Size

Sampling

Representativeness

Balance

(Tognini-Bonelli, *Corpus linguistics at work*, 2001:47-64)

English Corpora

- **The Brown Corpus (1964)**

1 million words (500 samples/2,000 words, written American English, texts published in the US in 1961

- **The Lancaster-Oslo/Bergen (LOB) Corpus (1978)** similar to the Brown corpus, British English, text from 1961 (compiled 1970-1978)

English Corpora

- **The London-Lund Corpus (LLC)**

200 samples, ~5000 words each, 1953-1987, spoken British English, transcribed.

- **The Frown Corpus**

Freiburg-Brown Corpus of American English **(1992)** 1990s analogue to the Brown corpus (1 million words, written American-English.

- **The FLOB Corpus**

Freiburg-LOB Corpus of British English, 1990s analogue to the LOB corpus (1 million words, written British English).

English Corpora

- **The British National Corpus (BNC)**

100 million-word, samples of written texts (90m words) and spoken language (10m words).

- **The International Corpus of English (ICE)**

500 samples (300 spoken, 200 written), ~2,000 words each, 1990 onwards, 20 national varieties of English (e.g. UK, India, Singapore, Australia, India, Jamaica)

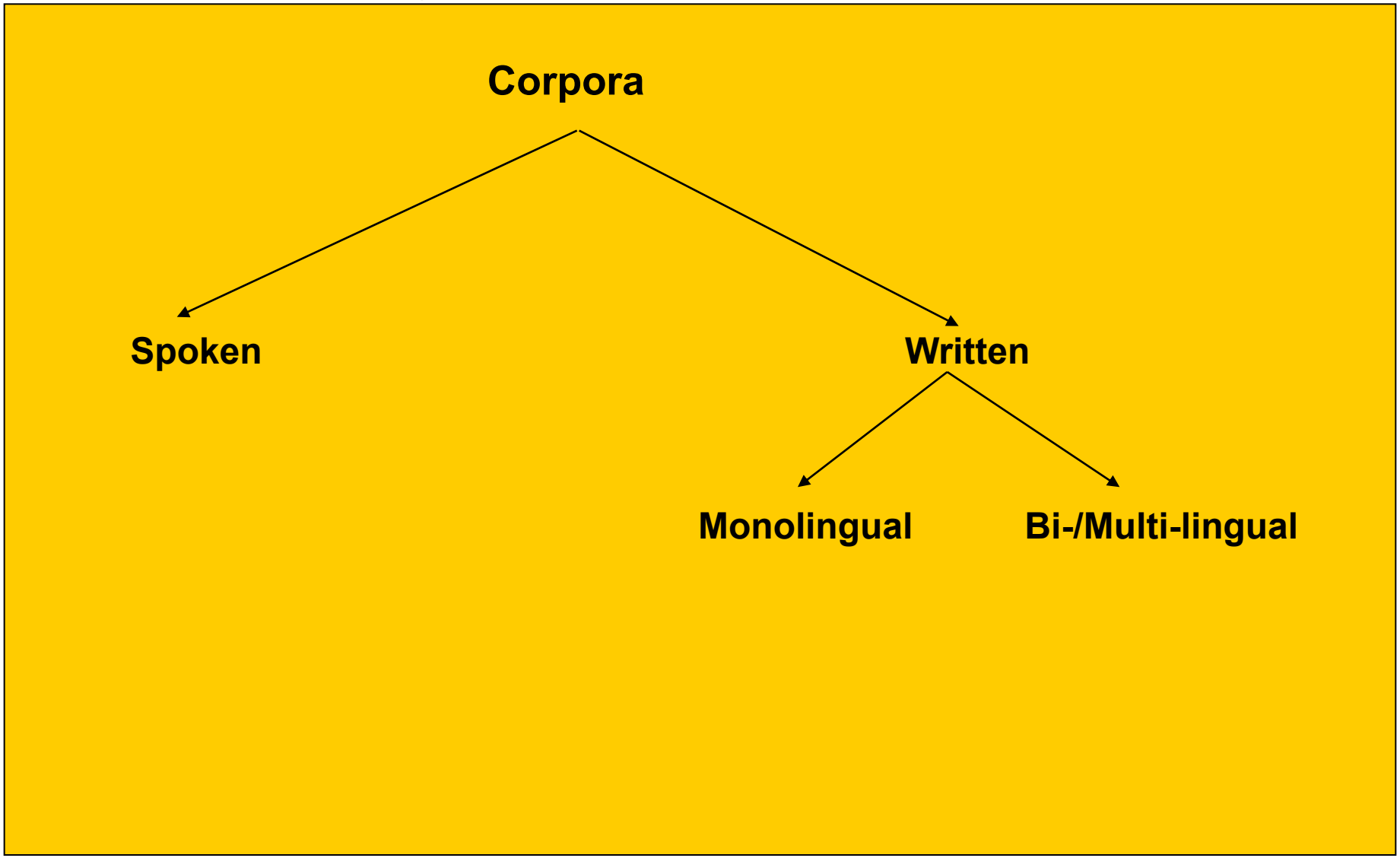
- **The BoE Corpus (The Bank of English Corpus)**

450M words, full texts, open, written and spoken, mainly US and UK

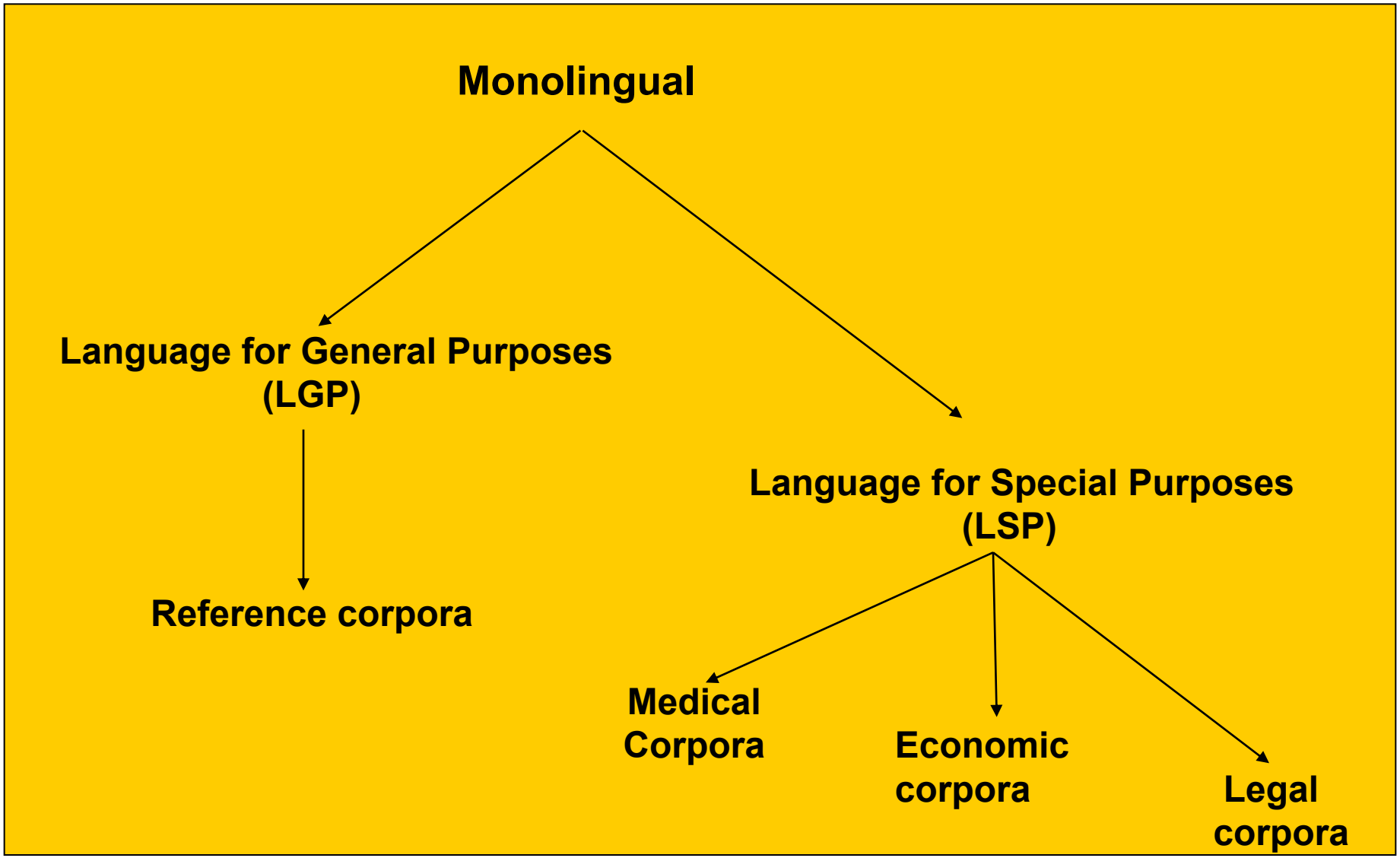
Types of corpora

- spoken vs. written
- monolingual vs. bi/multilingual
- parallel vs. comparable corpora (translation corpora)
- general language purpose vs. specialised language purpose
- diachronic vs. synchronic
- plain text vs. annotated (tagged) text

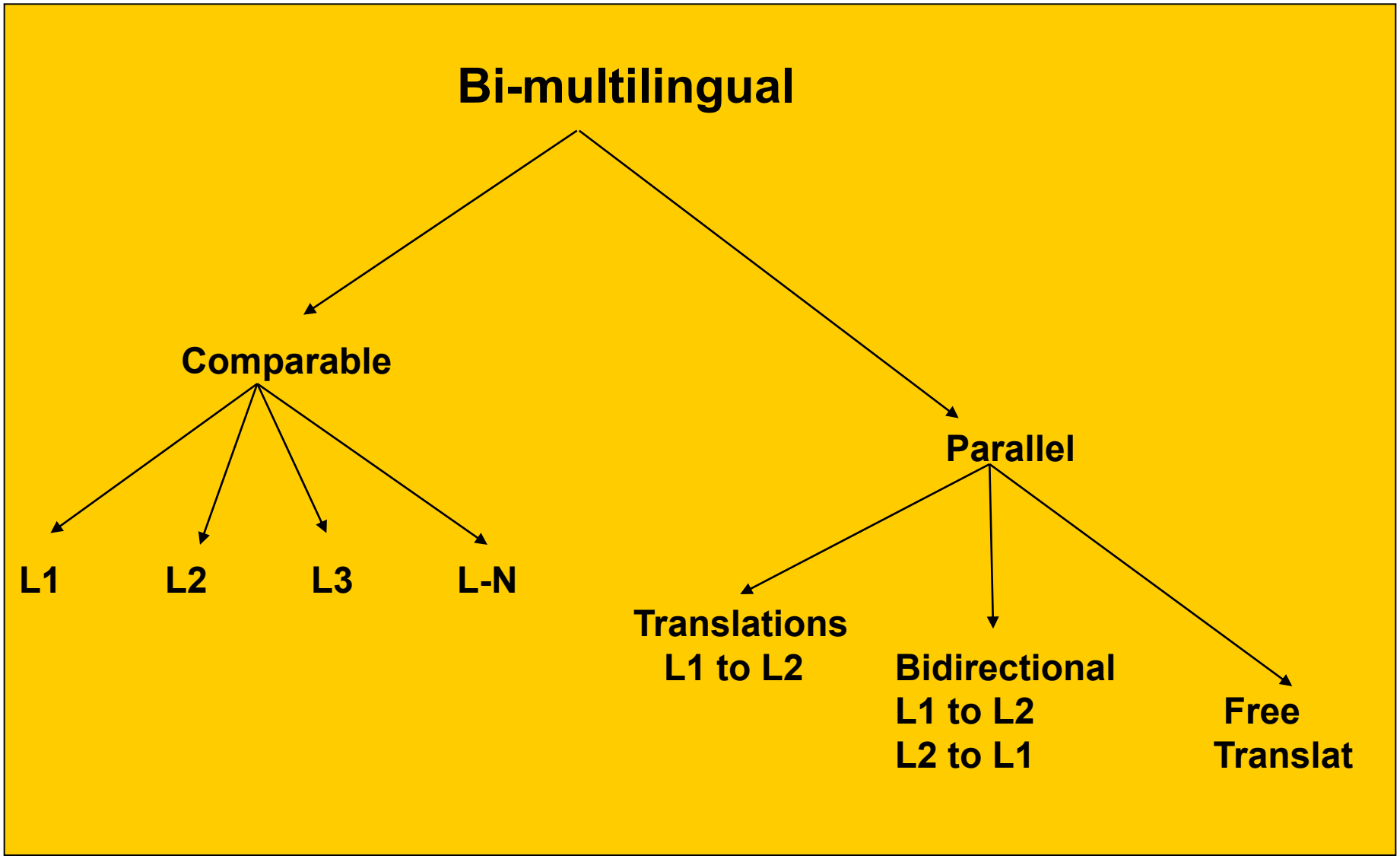
Types of corpora



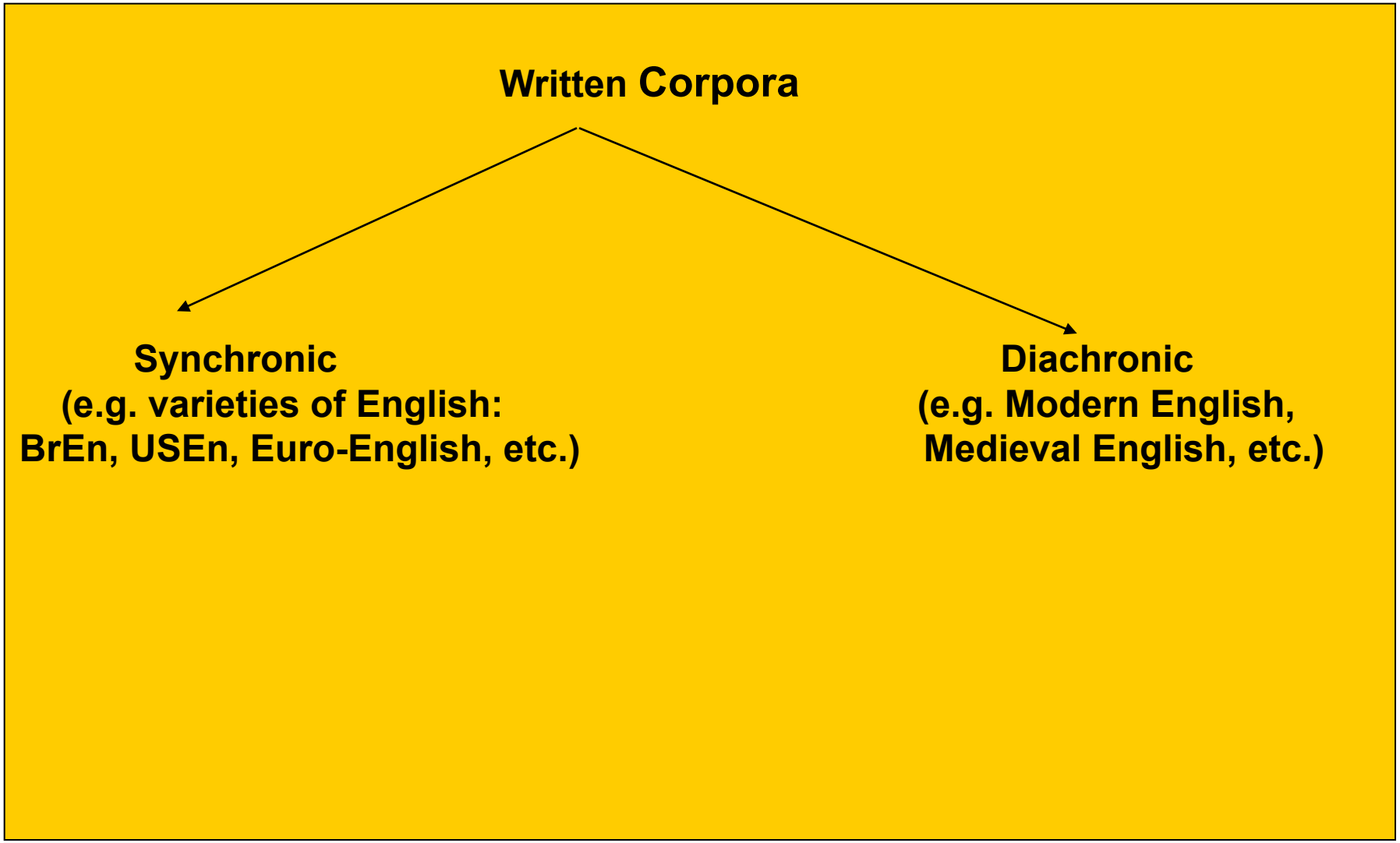
Types of corpora



Types of corpora



Types of corpora



Uses of Corpora

- ✓ **Lexicography / terminology**
- ✓ **Linguistics / computational linguistics**

  Dictionaries & grammars (*Collins Cobuild English Dictionary for Advanced Learners; Longman Grammar of Spoken and Written English*

  Critical Discourse Analysis

  - Study texts in social context

  - Analyze texts to show underlying ideological meanings and assumptions

  - Analyze texts to show how other meanings and ways of talking could have been used….and therefore the ideological implications of the ways that things were stated

- ✓ **Literary studies**
- ✓ **Translation practice and theory**
- ✓ **Language teaching / learning**

  ESL Teaching

  LSP Teaching (*exemplar texts*)

Lexicography / Terminology (wikipedia.org)

General **lexicography** focuses on the design, compilation, use and evaluation of general dictionaries, i.e. dictionaries that provide a description of the language in general use. Such a dictionary is usually called a general dictionary or LGP dictionary. Specialized lexicography focuses on the design, compilation, use and evaluation of specialized dictionaries, i.e. dictionaries that are devoted to a (relatively restricted) set of linguistic and factual elements of one or more specialist subject fields, e.g. legal lexicography. Such a dictionary is usually called a specialized dictionary or LSP dictionary.

Terminology, in its general sense, simply refers to the usage and study of terms, that is to say words and compound words generally used in specific contexts.

Terminology also refers to a more formal discipline which systematically studies of the *labelling or designating of concepts* particular to one or more subject fields or domains of human activity, through research and analysis of terms in context, for the purpose of documenting and promoting correct usage. This study can be limited to one language or can cover more than one language at the same time (*multilingual terminology*, *bilingual terminology*, and so forth).

Lexicography and corpora

- Corpus-based lexicography started in England
- Corpus provides authentic uses of language
- Extract samples (concordance) to identify different senses
- Word Frequency information
- Help identify collocation, set phrase
  – Collocation : *file … patent*, *move on*,
  – Set phrase : *night and day*, *black and white*
- Most English dictionaries are now corpus-based.
  Oxford, Collins, Longman, Cambridge, Macmillan, …

Linguistics and Corpora

- Research on empirical linguistics
- Study language use in various aspects
  - Verify linguistic theory, e.g. the explanation of definite description,
  - Lexical studies e.g. study near synonymous ‘little’ ‘small’
  - Sociolinguistics : compare the different of languages produced from different social groups (m/f)
  - Cultural study e.g. differences found in 2 comparable corpora (British/American) ….

Language Teaching / Learning and Corpora

- **Corpus-based** vs. **Corpus-driven**

"the term *corpus-based* is used to refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study" (Tognini-Bonelli, *Corpus linguistics at work*, 2001:65)

Language Teaching and Corpus-based approach

- Corpus based : use corpus as a resource
- Knowledge :
  - Know better about English

    answer specific questions of certain words, phrases, structures.
  - Know where the problems are

    error analysis on a learner corpus
  - Know what should be taught

    word frequency, comparing native/learner corpora

Language Teaching and Corpus-based approach

- References :
  - create better references
    dictionary, grammar book, textbooks
  - verify certain hypotheses about languages
    find support examples / counter examples
  - use a native corpus as a reference
    see whether it is possible
    which one is more natural

Language Teaching and Corpus-based approach

- Corpus based : use corpus as a resource
  Syllabus design :
  - Native corpora => what are actually used
  - Learner corpora => what are the problems
  - Find out which aspects should be given priority
  - Lexical syllabus = focus on frequency of occurrence
  - How many words the students should know? What are they?
  - Knowing 90% or 95% of the words?

Language Teaching and Corpus-driven approach

"In a *corpus-driven* approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence. The corpus, therefore, is seen as more than a repository of examples to back pre-existing theories or a probabilistic extension to an already well defined system. […] Examples are normally taken verbatim, in other words they are not adjusted in any way to fit the predefined categories of the analyst; recurrent patterns and frequency distributions are expected to form the basic evidence for linguistic categories; the absence of a pattern is considered potentially meaningful." (Tognini-Bonelli, *Corpus linguistics at work*, 2001:84)

Language Teaching and Corpus-driven approach

- Corpus driven
  - provides new paradigm of teaching/learning
  - students as a researcher
  - data driven learning
  - learn how to use concordance + corpora
  - extract generalization from data
  - Is it possible?

Corpus-based Translation

- Theoretical issues:

Descriptive Translation Studies: Toury, Baker, Laviosa, Teubert

- Creation of *parallel corpora* or *translation corpora*
- Alignment techniques:

Olivier Kraif – *Translational Compositionality* and *Maximal Resolution Alignment*

Corpus-based Translation

- Corpora as a resource for translation
- Parallel corpora / Translation memory
  - Provide examples of translation
  - TM software detect the most likely translation
- Native corpora
  - Help editing translation to be native-like
  - Help understanding difficult words/concepts

Corpus-based Translation

- Many experiments confirm that
  – Native corpora is useful for selecting the appropriate translation
  check whether that translation is possible;
  if > 1 translation choice, select the most occurrence
  – Native corpora help understanding the source text
- Translation school should teach students how to use corpora as a resource for solving translation problems.

Why to use a *corpus*?

- Intuition alone is not enough
  – Is "*starting*" always replaceable by "*beginning*"?
  – Is it only "*time*" that is "*immemorial*"?
  – "*think of*" vs. "*think about*"

- Native speaker intuition is unreliable
  – provides no information on frequency of occurrence
  – "*head*" => body part - Is this the most used sense?

- Help answering questions of usage easily
  – More than one character *is/are*
  – *Worth to* do / *worth* do*ing*

- Is it *sheer* a synonym of *pure, complete, utter* and *absolute*?

Text vs. Corpus
(Tognini-Bonelli 2001: 3)

TEXT	CORPUS
Read whole	Read fragmented
Read horizontally	Read vertically
Read for content	Read for formal patterning
Read as a unique event	Read for repeated events
Read as an individual act of will	Read as a sample of social practice
Coherent communicative event	Not a coherent communicative event

Text vs. Corpus

From time to time there is also the need for high quality information to support particular initiatives, such as the (successful) application for accreditation. Some progress has been made in recording data on the Polytechnic 's rooms and buildings, and on the teaching space requirements of individual courses. These data are analysed, along with the database on course details and students ' course and module registrations, using the methodology in DES Design Note 44. Ad hoc reports are an essential part of any system that aspires not merely to process data routinely but to permit management information to be creamed off the top.

N	Concordance		
13	ement system. They can choose whether to enter	data	themselves or to use the data preparation se
14	individuals is recorded. As well as student- related	data,	details on courses, on modules and their
15	student on the module is entered, for subsequent	data	processing by the registry. In addition
16	s detailed record- keeping and effortless access to	data.	All of this the system provides.
17	nd passed to the Registry who, together with the	data	preparation service, enter and verify approxi
18	considerable use of student management system	data	processing. The marksheet for each module
19	ons that can assist their work. Individual student	data	are available to assist counselling. Registry
20	tion. Some progress has been made in recording	data	on the Polytechnic 's rooms and buildings, a
21	onitors to allow the whole Committee to view such	data.	A detailed analysis of the performance of
22	space requirements of individual courses. These	data	are analysed, along with the database on c
23	easy to record in a computer (unless complicated	data	structures are used) and are even harder to

Corpus Linguistics : Some basic notions

- Concordance / Concordancer
- Collocation (Lexis)
- Colligation (Grammar)
- Semantic Preference (Semantics)
- Discourse Prosody (Pragmatics)

****

- Paradigmatic and Syntagmatic Dimensions
- Lexico-grammar approach
- Idiom principle vs. open-choice principle
- Phraseological tendency vs. terminological tendency
- Pattern (grammar)
- Extended units of meaning
- Cultural Keywords

Concordance / Concordancer

- ***Concordance***

  A term that signifies a list of a particular word or sequence of words in a context. The **concordance** is at the centre of **corpus linguistics**, because it gives access to many important language patterns in texts. **Concordances** of major works such as the Bible and Shakespeare have been available for many years. The computer has made **concordances** easy to compile.

  The computer-generated **concordances** can be very flexible; the context of a word can be selected on various criteria (for example counting the words on either side, or finding the sentence boundaries). Also, sets of examples can be ordered in various ways. See Sinclair 1991: Ch. 2; McEnery and Wilson 1996: Ch. 1; Collier 1994; Kaye 1990; Hockey and Martin 1988.

Concordance sample of *data* (BNC World Edition)

N	Concordance		
551	d proper nouns. The initial construction of the	data	structure is of little importance to the user
552	); the efficiency of representation of the	data	so that its particular features are succinctl
553	structure; the ease of alteration of the	data	structure (i.e. adding and deleting items);
554	Alternative data structures Looking at possible	data	structures for representing such a word list
555	pointer to the next word in the list. This	data	structure is extremely simple to implement
556	it is rarely performed. Alternative	data	structures Looking at possible data structu
557	the movement of the stylus across its surface.	Data	is collected in the form of x, y co-ordinates
558	nd ensuring that facilities are available for these	data	to be reported, analysed and evaluated. Ri
559	iding managers with easy access to high- quality	data	and ensuring that facilities are available for
560	er can make rapid comparisons between sets of	data.	This can be used to highlight changes fro
561	one location and containing a limited amount of	data	should be set up. This would give the staf
562	ed in the creation, updating and processing of a	data	base (the basic information to be stored in
563	mation contained on the payroll tape held by the	Data	Processing Branch. However, as the nee
564	Large sums of money could easily be spent on	data	collection and maintenance with a very limi
565	mplexity of drawing together accurate personnel	data	in a dispersed organisation — some
566	and centrally at TSB Group Central Executive.	Data	are captured partly through the computeriz
567	SB Group Central Executive. Where possible,	data	are collected direct from banking computer
568	tablishment levels, and relief requirements. The	data	are held on a branch-by-branch basis, and
569	ric indexation but the use of a computer enables	data	to be retrieved on a range of factors, either
570	ightforward and requires little effort to extract the	data.	Since titles, footnotes and other adjustm
571	ourses in extracting, analysing and interpreting	data	At each stage of development we have had

Collocation

- You shall know a word by the company it keeps (Firth 1957:179)
- We may use the term node to refer to an item whose *collocations* we are studying, and we may define a span as the number of lexical items on each side of a node that we consider relevant to that node. Items in the environment set by the span we will call *collocates*. (Sinclair 1966:415)
- *Collocates* are the words which occur in the neighbourhood of your search word (Scott 1999 WordSmith Help File).
- This a lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text. For example, PROVIDE frequently occurs with words which refer to valuable things which people need, such as *help* and *assistance*, *money*, *food* and *shelter*, and *information*. These are some of the frequent *collocates* of the verb. (Stubbs 2002: 24).

collocates …node …collocates

---------- span ----------

Concordance sample of *data*
(BNC World Edition) – collocations. Alphabetically sorted (R-1)

N				
	Concordance			
10289	and output to the supply rails. The RX	data	input	is clamped to the supply rails by diodes
10290	and a project to clear backlogs of registrations and	data	input	for borehole logs, with the intention of pr
10291	were required to update the PMIS. The ideal	data	input	document
10292	ndardised accounts automatically from accounting	data	input	by the analyst. An alternative is
10293	phase in direct proportion to the value of a 4- bit	data	input.	In the required circuit (figure,
10294	. This process is repeated for each source of	data	input.	The randomized input map data are the
10295	input/output lines are buffered from the computer 's	data	input/output	lines by IC5. This chip is an
10296	bit device with built-in Lithium battery. Its eight	data	input/output	lines are buffered from the compu
10297	circuit. If a 2-bit number is set up on	Data	inputs	D1 and D2 using switches S2 and S3,
10298	1 and D2 to avoid possible confusion later with the	data	inputs	D1, D2 etc. Following the
10299	been laid out like a form with clear headings and	data	inserted	in appropriate places. Whenever an
10300	the computer and it proved impossible to get at the	data	inside,	you could go to your back-up diskette
10301	re so unselective as to be innocuous; the Swedish	Data	Inspection	Board took the opposite view when
10302	lf to use by managers and policy makers as survey	data	instead	of as individual assessments. Spurio
10303	ok for remedies at classroom level from such global	data.	Instead,	it may be more appropriate to c
10304	Plato 's two-worlds theory of ideal forms and sense	data.	Instead,	they believed in the organic unity
10305	99 different vehicles by the American Highway Loss	Data	Institute.	Tougher driving tests start
10306	the end of 1993, and will include features such as	data	instruction	analysis, security and integrity, an
10307	Research Institute (IKI) would have access to the	data.	INTEGRAL	was given top ranking over thr
10308	evel architecture with abstractions and views, meta-	data	integrated	with operational data, short-term tr
10309	lip; has 28 years experience at the leading edge of	data	integration	technology, designing, producing,
10310	spatial units, one of the most intractable of all	data	integration	problems. A fuller description of th

Concordance sample of *data*
(BNC World Edition) – collocations. Alphabetically sorted (L-1)

N	Concordance				
9082	into institutional care. The wealth of	qualitative	data	collected in this research reveals a number of
9083	cost form which I 've and er,	qualitative	data	of uniform, not to mention any other, s
9084	mes of the education process. Although	qualitative	data,	which are far more subjective, are usually pr
9085	of their career at the Bar, and also	qualitative	data	such as attitudes towards careers for women
9086	ysis of a large amount of quantitative and	qualitative	data.	It also requires presentation in a manner acc
9087	Secondary analysis of data	Qualitative	data	Multidisciplinary modular courses require con
9088	've qualitative field review forms, yeah	qualitative	data	review forms yeah qualitative research approva
9089	ot be performed by manual means with poor	quality	data	&equo; (Openshaw 1980 : 289). Heywood
9090	state that the estimates are based on high	quality	data	supplied by The Scotch Whisky Association '
9091	rial. Where significant new or better	quality	data	have become available, revision of some of th
9092	tronomical Satellite (IRAS) is turning in high-	quality	data	in such profusion that astronomers are having
9093	fficulty was always to obtain sufficiently high-	quality	data	within a defined area. In 1974, John Miche
9094	emergency it is of course vital to have high-	quality	data	on the distribution of population and resources
9095	roviding managers with easy access to high-	quality	data	and ensuring that facilities are available for the
9096	Statoil. &bquo; The ability to derive	quantative	data	from the core within hours of coring played a
9097	serve the requirements of objective and	quantifiable	data	in the variable analytic format. The early
9098	d reinforcement for the client. Third,	quantifiable	data	is the cornerstone of applied clinical research.
9099	factors. Some specify the collection of	quantifiable	data,	others use more subjective qualitative forms
9100	te large amounts of quantified or easily	quantifiable	data.	The primary purpose of these data is to
9101	there is the framework of a sub- regional,	quantified	data	base on environmental decline. Also referenc
9102	can never be reliable enough to use as	quantitative	data	defining the presence or absence of consciou
9103	es generate appropriate consideration.	Quantitative	data	By their very nature modular courses generat
9104	le line on their techniques; and, while	quantitative	data	are conventionally available to other researche
9105	ry scale these studies aimed to gather	quantitative	data	on use. In spite of the breadth of the sample

Colligation

- Colligation can be defined as ‘the grammatical company a word keeps and the position it prefers’: in other words, a word’s *colligations* describe what it typically does grammatically (Hoey 2000:234)

- knowledge of a collocation, if it is to be used appropriately, necessarily involves knowledge of the *patterns* or *colligations* in which that collocation can occur acceptably (Hargreaves 2000:214).

Concordance sample of *give* (BNC World Edition) - colligations

N	Concordance				
4194	o through them all this morning. Let me	just	give	you	a list of some of the things that, tha
4195	r own computer which is I B M compatible we	just	give	them	the disk because it 's programmed that t
4196	using either of those ways. Or you can	just	give	yourself	some bullet points because you do n'
4197	normal with us at the moment. I might	just	give	him	another run — if only to keep the s
4198	the clothes line , stand back a few inches,	just	give	it	an even spray,
4199	e morning. Right.	Just	give	it	to the lassie and she 'll put you through
4200	Edwardian town house in London expecting to	just	give	it	a lick of paint. But shortly after set
4201	you f for this er equation but they	just	give	it	to you. And they give you all the
4202	h. Just give me a just give me a ring. yeah.	Just	give	me	a just give me a ring. B
4203	sort of erm instrument thing you know you can	just	give	that.	Truly sir, tru
4204	listed T C I P. Right, if you	just	give	the	command Plot T C Plot space T C then
4205	much more important. Okay. Let me	just	give	you	an idea, just that 's something else we d
4206	Are you listening? A la carte then?	Just	give	me	a. In the erm the
4207	the project that is, you know, you ca n't	just	give	dole	out to white people and refuse it to black
4208	control. And the before nine guarantees,	just	give	them	the time factor again. Er, that enabl
4209	ight, what we 're going to do now, is	just	give	you	your er, pieces of paper back so you 'll
4210	Er … forgot who this is … so let's	just	give	him	the popular name &bquo; Local &equo;,

Semantic Preference

Semantic preference is the relation, not between individual words, but between a lemma of word-form and a set of semantically related words, and often it is not difficult to find semantic label for the set. […] [An] example is the word-form *large*, which often co-occurs with words for "quantities and sizes". (Stubbs 2002: 65)

Semantic or Discourse Prosody

A discourse prosody is a feature which extends over more than one unit in a linear string. […] Discourse prosodies express speaker attitude (Stubbs 2002: 65)

'the consistent aura of meaning with which a form is imbued by its collocates' … prosodies based on very frequent forms can bifurcate into 'good' and 'bad', using a grammatical principle like transitivity in order to do so. For example, where *build up* is used transitively, with a human subject, the form of the prosody is uniformly good … Where things or forces, such as *cholesterol, toxins*, and *armaments build up* intransitively, of their own account, they are uniformly bad. (Louw 1993:171)

Concordances of *build up*

N	Concordance
22	for a week, and using a trampoline at home to build up my fitness, and I 've been keeping a clos
23	having fun is quite a structured exercise. They build up one structure which breaks down and flow
24	egies as part of a positive parenting approach. Building up parental confidence in these technique
25	id out on different colours and a lot of time to build up a collection. I did not start on
26	rn blind; it sleeps and eats a great deal, gradually building up its size and strength. It is important
27	Something approaching a personal crisis had been building up since Nietzsche 's return from the unfor
28	and, for this to happen, a listener must be building up an analysis of both aspects while proc
29	And so we still have a common struggle and to build up our international solidarity. In conclusion
30	every month, making its a valuable part- work which builds up into a library on development, a handy s
31	better coordination were a matter of the patient building up of contacts, although the improved par
32	sually need only a light pruning, (a) , to build up strong main stems and develop lateral sid
33	Slowly, like a black storm cloud that builds up ominously on a distant horizon, the seco
34	man who made a personal fortune of £15m building up and selling companies like Kwik-Fit, B
35	to show different sides of a thought or feeling, building up the sense- impressions that cluster aro
36	off. I 'll take a shorter run up and build up to it gradually. Male speaker
37	and the book consisted of a series of dialogues, building up with phrases rather than individual word
38	the end of the 19th century, a crisis had been building up over the names of organic compounds.
39	courage students with a wide spectrum abilities to build up a fluency and accuracy in spoken and writ
40	the the Euros and well to a certain extent we 're building up already for any any editor content.
41	will take place, once a reasonable level of bacteria builds up. More on this later. The air
42	EEC institutions, had devised a plan that aimed to build up cultural, defence and foreign policy co-op

Semantic or Discourse Prosody

N	Concordance				
4157	Global Warming Climate change	could	cause	crisis in China A team of scientists	fro
4158	over the business and eventual retirement	could	cause	problems but this will be discussed	later.
4159	's operating system is quote robust, you	could	cause	the Z88 to enter an undefined state	if you
4160	nitial expressions of fear that this practice	could	cause	problems, it does not in fact seem	to ha
4161	lays in implementing the raft of legislation	could	cause	real problems. On the single market,	th
4162	ave agreed on the proper procedure. This	could	cause	the offended party to feel they have	been
4163	s products offered by other intermediaries	could	cause	an increase in the early termination,	or &
4164	r he shot himself. &bquo; Hardly anything	could	cause	a more widespread and painful sensation	
4165	ndent on Sunday. Although the amounts	could	cause	the deaths of up to 500,000 people,	the c
4166	against pesticide traces in drinking water	could	cause	serious difficulties for the water industry.	
4167	a mile for cars and 4.5 pence for lorries	could	cause	one tenth of motorway traffic to divert	to
4168	gns will not be allowed if they	could	cause	a traffic hazard. Woman	
4169	le as initially enunciated without limitation	could	cause	very serious practical difficulties	of admin
4170	t five hours after the blast. Contamination	could	cause	lung problems, similar to bronchitis,	and
4171	pack of the Galileo space probe, said it	could	cause	a nuclear accident worse than Chernobyl	
4172	a 10 per cent decrease in the ozone layer	could	cause	300,000 cases of skin cancer a year	worl
4173	ut since medical evidence proved that this	could	cause	severe metacarpal damage they have	be
4174	n cyclic variation in the list sequence that	could	cause	bias. For example, if we pick every	tent
4175	Mm. that 's a T S R.	Could	cause	unexpected results. T S R. Could	caus
4176	are not sure whether rising temperatures	could	cause	a catastrophic rise in sea levels.	Sea
4177	eois backlash, recognising the damage it	could	cause	to the already frail economy. With	Zhao

Semantic or Discourse Prosody

N	Concordance				
2756	be relatively easy by comparison. The	British	provide	a stable government and a way of life re
2757	uting 's long-term strategy to become a	broader	provider	of the elements of the network computi
2758	otland Please find some copies of the	brochure	providing	information on the dates and venues f
2759	rmation to local communities Site	brochures	provide	environmental information to local com
2760	everal of the better equipped European	brothels	provided	chambers decorated like railway carri
2761	d setting of raw flintstones at Johnson	Brothers	provides	a nostalgic view of two remaining bottl
2762	n take it on head- to- head. Unfortunately	Brown	provides	no numbers estimating the size of this
2763	chart or network. The preparation of a	budget	provides	a measure against which actual perfor
2764	ke place. If carried out correctly,	budget	provides	an effective way of quantifying the perf
2765	or current tenants. The 1993–94	budget	provides	for a £2m contribution to this b
2766	as the project came to a close the total	budget	provided	in the grant for support workers began
2767	. Is this something new for the new	budget	provide	more training for carers and are we goin
2768	. (ii) These	budgets	provide	a basis for responsibility accounting. (
2769	O THE CAR BOOT. A MOTORISED	BUGGY	PROVIDES	INDEPENDENT MOBILITY. BACK
2770	and carry loads. A motorized scooter or	buggy	provides	more mobility with less effort. In all c
2771	ld Company building. The new station	building	provided	facilities of waiting rooms, toilets and
2772	. The major halls in the Richmond	building	provide	excellent venues, particularly for large
2773	Royal Pavilion. The elegant Regency	buildings	provide	comfortable hotel accommodation to su
2774	to allow you some introductions. These	bulk	providers	of work, however, are not without th
2775	RAC These three sources of work are	bulk	providers	with well- established lawyers in most
2776	to seek help voluntarily •		provide	advice and any other type of help need

Phraseological tendency vs. Terminological tendency

- Sinclair puts phraseology at the heart of language description, arguing that the tendency of words to occur in preferred sequences has three important consequences which offer a challenge to current views about language:

1. There is no distinction between pattern and meaning;
2. Language has two principles of organisation: the idiom principle and the open-choice principle;
3. There is no distinction between lexis and grammar.

1. There is no distinction between pattern and meaning

- Different meanings for a word tend to be used in different grammatical patterns:
  - "Maintain something"
  - "Maintain that something is true"
  - "Maintain something at a level"
- Different grammatical patterns tend to collect words with similar meanings
  - VERB one's way (in)to: *bribe, bully, cheat, fiddle, hustle, insinuate, trick, wrangle….*

2. Language has two principles of organization: the idiom principle & the open-choice principle

The *open-choice principle* "is a way of seeing language text as the result of a very large number of complex choices. […] This is probably the normal way of seeing and describing language. It is often called a 'slot-and-filler' model, envisaging texts as a series of slots which have to be filled from a lexicon which satisfies local restraints." (Sinclair 1991: 109)

These restraints are mainly *grammatical*.

2. Language has two principles of organization: the idiom principle & the open-choice principle

But words “do not occur at random in a text”

“The choice of one word affects the choice of others in its vicinity. Collocation is one of the patterns of mutual choice, and idiom is another. The name given to this principle of organization [*of language*] is the ***idiom principle***.” (Sinclair 1991: 173)

In other words, “the language user has available to him a large number of preconstructed or semi-preconstructed phrases that constitute single choices, even though they appear to be analysable into segments”. (Sinclair, quoted in Partington 1998: 19)

2. Language has two principles of organization: the idiom principle & the open-choice principle

- Idioms:

to get a frog in one's throat vs. **to get an ugly frog in one's throat*

- Examples of idiomaticity:

Of course (= *insofar as*)

- Phrases allowing internal lexical variation:

In some cases / in some instances / set x on fire / set fire to x

- Phrases allowing internal syntactic variation:

It's not in his nature to …

  - The verb tense can vary (*was*) or a modal may be introduced;
  - The negative *not* can be substituted with another negative (*hardly*)
  - The possessive *his* can be substituted with *my, your, 's*

- Phrases allowing some variation in word order

  to recriminate is not in his nature vs. *it is not in the nature of an academic to …*

- Words and phrases showing a tendency to co-occur with certain grammatical choices

  set about (=inaugurate)

Irreversible collocations
cash and carry

N	Concordance
20	y, SHV, and is based in Manchester, runs cash and carry operations for independent
21	o become National Account Manager in the Cash and Carry and Wholesale Sector fro
22	the Contemporary Art Society Market, the Cash and Carry Art — the Sainsbur
23	ed good value occasional type furniture the cash and carry the kind of things that you
24	hardly ever go. My father will go to the cash and carry every week, and perhaps h
25	carry anyway. You 're going to the cash and carry, so. I 'm going to
26	carry, so. I 'm going to the cash and carry, yeah. I was actuall
27	Yeah. things erm in the cash and carry promotions. So she 's got
28	. Well I 'm going to the cash and carry . Well I 'm going to the cas
29	. I 've got I 'm going to the cash and carry anyway. You 're going to
30	ards operates throughout Ireland, and their cash and carry operation has two fully equi
31	under 's generic brands and to cash and carry 's. A combination of acqui
32	laiming around five percent of the overall UK cash and carry trade. According to comm

Irreversible collocations
bread and butter

N	Concordance							
71	Corp Unix variant it markets &bquo; is still our	bread	and	butter	business. &equo; The Intera
72	o it blindly, just hoping that we could earn our	bread	and	butter.	We had no idea we would a
73	ssel and watercress soup Lamb in puff pastry	Bread	and	butter	pudding with whisky and hon
74	auses; it is to provide the jam on the plain	bread	and	butter	the Treasury says it will cont
75	basis. Typically these systems involve "	bread	and	butter	" operations within an e
76	case. &equo; Despite Sparcbook, Tadpole 's	bread	and	butter	business remains its VME a
77	rm club acts and to undertake the musician 's	bread	and	butter	work playing in theatre band
78	RECORD If League football is a club 's	bread	and	butter,	then a good Cup run must
79	rich and &bquo; rags to riches was filmdom 's	bread	and	butter	&equo;. The content of the fi
80	netting? That 's no contest. That 's	bread	and	butter	that is. Besides, I never
81	mint. Anton Mosimann 's	Bread	and	butter	pudding Serves 6–8
82	Ltd. The VLIW stuff aside, Equator 's	bread	and	butter	comes from designing custo
83	in beef dripping, airy and crisp, plus sliced	bread	and	butter	and a pot of tea-bag tea. Wa
84	, Marge, have a nice bit of fish with some	bread	and	butter.	&equo; But he ended up ea
85	st meal for six days, before being given some	bread	and	butter	and cigarettes and being advi
86	couple of weeks ago Do you want some	bread	and	butter	with jam Charlotte? Yeah
87	still have to eat, and our ships do a steady	bread	and	butter	trade in produce from the Co

Irreversible collocations
salt and pepper

N	Concordance						
25	a bowl and then add potatoes, onion, flour,	salt	and	pepper.	Mix well. Heat a		
26	Slowly add the vinegar and stir in the sugar,	salt	and	pepper.	Split the baked potatoes an		
27	Add the rice, cooked chick peas, herbs,	salt	and	pepper	to the onion mixture. Arrange		
28	bottles. Two gallons are used daily.	Salt	and	pepper	sit on the tables in old jam-jar		
29	tatoes through a sieve then beat in the butter,	salt	and	pepper	until the mixture looks cream		
30	1 red pepper soy sauce	salt	and	pepper	1 tbsp oil Chop		
31	soy sauce dash Worcester sauce	salt	and	pepper	small tin tomatoes		
32	Whisk the eggs with a little water and some	salt	and	pepper	and pour over the vegetables.		
33	ped fresh marjoram or oregano or 1 teaspoon	salt	and	pepper	recipe ends here		
34	. Can I take the	salt	and	pepper	through or do you need it		
35	Finally, add the rest of the stock and the	salt	and	pepper.	If the rice is still not c		
36	fat cottage cheese. 8 tomatoes	salt	and	pepper	6 oz Shape low fat		
37	l sauce is mushy and thickened. Season with	salt	and	pepper.	Meanwhile cook pasta in ple		
38	lender or food processor and season well with	salt	and	pepper.	Blend for a few minutes until		
39	hallot and bouquet garni and season well with	salt	and	pepper.	Cover and chill for 1 hour		
40	in the milk to make a batter. Season with	salt	and	pepper.	Heat the butter in a f		
41	eese and the parsley and season to taste with	salt	and	pepper.	Do n't forget So I		

Irreversible collocations
black and white

N	Concordance					
19	r on-screen may not appear emphasized on	a	black	and	white	printout. So unless you 're p
20	BG in 1979 but still without a guidebook.	A	black	and	white	leaflet is available for the pub
21	ects, and they will therefore not show up on	a	black	and	white	viewfinder. Flesh
22	DeskJet 550C can, of course, be used as	a	black	and	white	printer, and is a solid, well b
23	r bedroom with its mattress on the floor and	a	black	and	white	duvet looks spacious, vergin
24	a ha. So it 'll probably be	a	black	and	white	So it 'll probably be a bla
25	we 'd get a television for him and we had	a	black	and	white	television Would you remem
26	re he took his glasses off and he was like	a	black	and	white	minstrel. Wait till the bl
27	helds. It runs an eight- bit 6502 chip,	a	black	and	white	screen that is not the easies
28	heck to go with black and white shoes and	a	black	and	white	striped shirt. Ballesteros sh
29	ather in my cap. &equo; His cap was	a	black	and	white	check to go with black and
30	bits as we get them. Would	a	black	and	white	photocopy of a map I
31	erviews when suddenly there was a pause.	A	black	and	white	photo of a young man appea
32	y page printer can, of course, only produce	a	black	and	white	image which can incorporate
33	picture of a man in uniform. It was	a	black	and	white	photograph which had been
34	, artist, size, date it was stolen and	a	black	and	white	illustration. Most of the item
35	as a very interesting sign in the shape of	a	black	and	white	pig hanging at the entrance.
36	I figured that rather than have a tape and	a	black	and	white	photo, I would have a CD th

Irreversible collocations
white and black

N	Concordance										
1	has refused the opportunity. New Level, a	white	and	black	son of Murlens Slippy, took part in						
2	We 're half a mile long and young and old and	white	and	black	and girl and boy, looking for a mon						
3	look better with it I reckon, cos It will be	white	and	black	(or blue?), rather than just						
4	ecome open so that there can be reciprocity between	white	and	black	society. The child 's psychic structur						
5	s — a pressure point for racial tension between	white	and	black,	and between African and Caribbean.						
6	. Even at this early stage of contact between	white	and	black	in Britain, it is clear that she wo						
7	6 Miscegenation (sex relationship between	white	and	black	race) is forbidden. 7						
8	; though Aggrey firmly believed in partnership between	White	and	Black.	Looking back, Nkrumah pronounc						
9	black &equo; and &bquo; white &equo;, but between	white	and	black	men (West wood, 1990), white						
10	en the main reason for expulsion was common to both	white	and	black	pupils, the latter were more likely to						
11	lack of data on the level of infection in both	white	and	black	people, according to Nicky Padayac						
12	The point is that the fact of empire affects both	white	and	black	communities. Much of Britain 's pros						
13	few months to four years. The majority, both	white	and	black,	come from working-class backgroun						
14	f political violence in which victims have included both	white	and	black	children. The president said 18						
15	Central African Federation &equo;. It is what both	white	and	black	in this country have met in dealing wi						

There is no distinction between lexis and grammar

- To know a word is to know how to use it
- Certain grammar attracts certain words
- Grammatical words like *a* and *the* are often used in phrases rather than being used independently
  - *A free hand* vs. *her free hand*
  - *Hurt his leg* vs. *hit someone in the leg*
  - *Turn her face* vs. *a slap in the face*