

Introduction to Corpus Linguistics

What is a corpus?

- A corpus is a collection of **naturally-occurring** language **texts**, chosen to characterize a **state or variety** of a language. In modern computational linguistics, a corpus typically contains **many millions of words**: this is because it is recognised that the creativity of natural language leads to such immense variety of expression that it is difficult to isolate the recurrent patterns that are the clues to the lexical structure of the language Sinclair (1991 :171) .

What is corpus linguistics?

(Teubert & Krishnamurthy 2007)

- The linguists are not in charge of the language; the discourse community is (ibid.: 9)
- “The discourse community establishes the conventions for what is acceptable and what is not” (ibid.: 9)

What is corpus linguistics?

- “Corpus linguistics is concerned with meaning, with symbolic content. People are not interested in grammatical constructions; they want to know the meaning of what has been said.” (ibid.: 9)

What is corpus linguistics?

- “What sets corpus linguistics apart from cognitive linguistics is that it looks at language from a social, not a psychological perspective. Language is verbal communication between people, is the discourse of what is actually being said (written) and listened to (read).” (ibid.: 9)

What is corpus linguistics?

- “Corpus linguistics is bottom-up ... accommodate the full evidence of the corpus. It analyses the evidence with the aim of finding probabilities, trends, patterns, co-occurrences of elements, features or groupings of features” (ibid.: 6) that form units of meaning
- “the starting point is always the corpus, real language data” (ibid.: 6)

What is corpus linguistics?

- “Corpus linguistics uses frequency to arrive at generalisations. Statistical significance makes us aware of connections that we would not see otherwise. The generalisations that corpus linguistics arrives at are not interpreted as laws or rules, but as plausible ways to group similar things together.” (ibid.: 9)

What is corpus linguistics?

- “Corpus linguistics can also make specific claims concerning unique events of language phenomena by showing in which aspects this event differs from all other occurrences of the same type of phenomenon.” (ibid.: 9)

History of corpus linguistics (1)

Late 19th century

The Oxford English Dictionary, compiled by means of an enormous number of slips collected containing authentic examples of language in use

Late 1950s and early 1960s

- the beginning of proper corpus linguistics (Tognini-Bonelli 2001: 52)

History of corpus linguistics (2)

1959

- Randolph Quirk announced his plan to start a Survey of English Usage of both written and spoken English
- Not computerized because 50% spoken language
- *A Comprehensive Grammar of the English Language* (Quirk, Greenbaum, Leech, & Svartvik 1985)



The input of new technologies

The computer

- To assemble corpora (large amounts of data) from the Web, scan electronic databases on CD-ROM or connect to a database by remote access
- To store large amounts of information
- A very fast tool to process and systematise a quantity of information in real time (Tognini-Bonelli 2001: 5-6)

Corpus analysis software

- Software packages such as *WordSmith Tools* (Scott, 1999)
- The software “selects, sorts, matches, counts and calculates” (Hunston and Francis, 2000: 15)

Three stages of the computer corpus in linguistics work (1)

1. As a tool to process, in real time, a quantity of information
2. A distinctive and enhanced methodology of enquiry into language - to provide abundant new evidence in a speedy and systematic way

Three stages of the computer corpus in linguistics work (2)

3. Corpus linguistics is a domain of research; “a new philosophical approach to linguistic enquiry” (Tognini-Bonelli 2001: 1); re-unites the activities of data gathering and theorising which lead to a qualitative change in our understanding of language (Halliday 1993: 24)

Authentic language use

- “in the final analysis if linguistics is not about language as it is actually being spoken and written by human beings, then it is about nothing at all” (Trudgill 1996: xi)
- Corpus linguistics is the study of language through observation of language evidence in corpora. It differs from traditional linguistics in its insistence on the systematic study of authentic examples of language in use (Tognini-Bonelli 2001: 1).

The contextual theory of meaning (Firth 1957)

J.R. Firth (1880-1960) died before the advent of computers and electronic corpora, but “laid the theoretical foundation of a contextual theory of meaning which is central to our present-day view of corpus work” (Tognini-Bonelli 2001: 157).

Assumption of the contextual theory of meaning

“We must take our facts from speech sequences, verbally complete in themselves operating in contexts of situation which are typical, recurrent, and repeatedly observable. Such contexts of situation should themselves be placed in categories of some sort, sociological and linguistic, within the wider context of culture.” (Firth 1957: 35)

Applications of the contextual theory of meaning

- “Speech events have to be apprehended in their contexts, as shaped by the creative acts of speaking persons.” (Firth 1957:193)
- Firth’s (1957) contextual theory of meaning can be applied to:
 - the analysis of a text: language as function in context
 - the analysis of a corpus as a corpus contains texts

Why to use a *corpus*?

- **Dictionary explanation is not accurate**

- *Is 'place' mostly used to refer to the 'physical environment' as defined in the dictionary. Sinclair (2003) denied this point and found it is most frequently used in the phrase 'take place'.*

- **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 also unreliable**

- provides no information on frequency of occurrence
- “*head*” => body part - Is this the most used sense?



The Word Counter

- <http://www.youtube.com/watch?v=ixw-XyycGdU>

How to Read a 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
Read as a Coherent communicative event	Not a coherent communicative event

How to read a corpus

- 1. Read fragmented and vertical: **Concordance**
Concordance is 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. The computer has made **concordances** easy to compile.

Concordance / Concordancer

■ KWIC

- KWIC is an acronym for Key Word In Context, the most common format for concordance lines.

- A KWIC index is formed by sorting and aligning the words within a corpus search either in alphabetical order or in frequency order.

■ ***Concordancers:*** online concordancers; Softwares, like WordSmith Tools

<http://www.americancorpus.org/>

Text vs. Corpus

[-5 key word +5]

N Concordance

1 and continue to work with the definitions <w PRP>throughout the essay, relating the different definitions
2 The Dionysiac cult was established <w PRP>throughout Greece and a symbolic reconciliation with
3 the dispensers. yet he did not utter a groan <w PRP>throughout the operation. Afterwards she found herself
4 Throughout
5 to make the king their puppet continued <w PRP>throughout his minority. The queen mother remarried
6 and (iii) the trees that are shade-bearers <w PRP>throughout life and are incapable of rapid growth in the
7 Available worldwide THIS TEXT IS AVAILABLE THROUGHOUT THE WORLD only as part of the
8 this situation prevails in all industries <w PRP>throughout the economy (that is, price equals
9 of the most significant features of their lives <w PRP>throughout the disability career. Achieving this
10 Trades Union of February 1834. <s n="13"><w PRP>Throughout his radical and union career Doherty
11 Type C may have been present <w AV0>throughout — from the early to the late-fourth
12 11 Mediation Attempts <s n="1204"><w PRP>Throughout the war, there were constant efforts to
13 the Good News of Christ's love spreading <w PRP>throughout the world. The present imbalance in the
14 a feeling of continued need for defecation <w PRP>throughout the distension period. Only one subject
15 on account of the high-grade circuitry used <w AV0>throughout. The combo offers the ubiquitous passive

Types of corpora

- Specialized corpus
- General corpus
- Multilingual corpora
- Comparable corpora
- Parallel corpora
- Free-translation corpus
- Learner corpus
- Pedagogic corpus
- Historical or diachronic corpus
- The Internet as corpus

(see, Hunston 2002: 14-16; Tognini-Bonelli 2001: 6-9)

Uses of corpora

- the tracking of changes in the English language
- the production of dictionaries and other reference materials
- the development of aids to translation
- language teaching materials
- the investigation of ideologies and cultural assumptions
- the study of all aspects of linguistic behaviour, including vocabulary, grammar and pragmatics
- the study of register variation
- natural language processing

US and British English corpora

All of these corpora spanning 60 years are based on written texts, and have used the same design criteria to allow comparisons to be made across the two varieties of English and across time.

US English

Brown (1961)

Frown (1991)

British English

BLOB (1931)

LOB (1961)

FLOB (1991)

Brown Corpus (1961) (1)

- A computerized corpus of US English
- “a standard sample of present-day edited American English, for use with digital computers” (Francis and Kucera 1979)

Major 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)

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

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

Monitor Corpora

The world's two largest corpora are in the UK:

Bank of English – approx. 500 m words

The Collins Wordbanks *Online* English corpus: 56 million words of contemporary written and spoken text. (<http://www.collins.co.uk/corpus/CorpusSearch.aspx>)

British National Corpus (BNC) – 100 m words

(<http://sara.natcorp.ox.ac.uk/lookup.html>)

British National Corpus (BNC)

- The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written. The latest edition is the *BNC XML Edition*, released in 2007.
- <http://www.natcorp.ox.ac.uk/>

BNC written

- The **written part** of the BNC (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text.

BNC Written (1)

- “The BNC was designed to characterise the state of contemporary British English in its various social and generic uses.” (Aston & Burnard, 1998: 28)
 - Imaginative 20%
 - Arts 8%
 - Belief & thought 4%
 - Commerce & finance 8%
 - Leisure 11%
 - Natural & pure science 4%
 - Applied science 11%
 - Social science 15%
 - World affairs 14%
 - Unclassified 2%

(p. 29)

BNC Written (2)

- Book 46%
- Periodical 36%
- Miscellaneous published 6%
- Miscellaneous unpublished 7%
- To-be-spoken 1%
- Unclassified 2%

(Aston & Burnard, 1998: 30)

BNC Spoken

- The **spoken part** (10%) includes a large amount of unscripted informal conversation, recorded by volunteers selected from different age, region and social classes in a demographically balanced way, together with spoken language collected in all kinds of different contexts, ranging from formal business or government meetings to radio shows and phone-ins.

The Cobuild project: Bank of English

- COLLINS Birmingham University International Language Database, 1980-1986
- The use of the computer plays “a clerical role in lexicography” (Sinclair 1991: p. 2)
- A huge database of annotated examples of language use was assembled
- A substantial dictionary edited from the database: *Collins Cobuild Dictionary* (Sinclair et al 1987)



American National Corpus

- The ANC is being developed to have, for American English, the kind of linguistic documentation that exists for British English in the British National Corpus.
- The goal for the ANC is to parallel the general structure of the BNC, while adding genres like blogging and instant messaging that did not exist when the BNC was created.

Online Resources

- 1. PolyU Language Bank
[The PolyU Language Bank](#)
- 2. Mark Davis's website
<http://corpus.byu.edu>
- 3. David Lee's website:
<http://tinyurl.com/r7zubf>
- 4. Sketch Engine
<http://ca.sketchengine.co.uk/login/>

Classification of Corpora (Mode)

Corpora

```
graph TD; A[Corpora] --> B[Spoken]; A --> C[Written]; C --> D[Monolingual]; C --> E[Bi-/Multi-lingual];
```

Spoken

Synchronous: online chatting;
online conferencing; instant
messaging

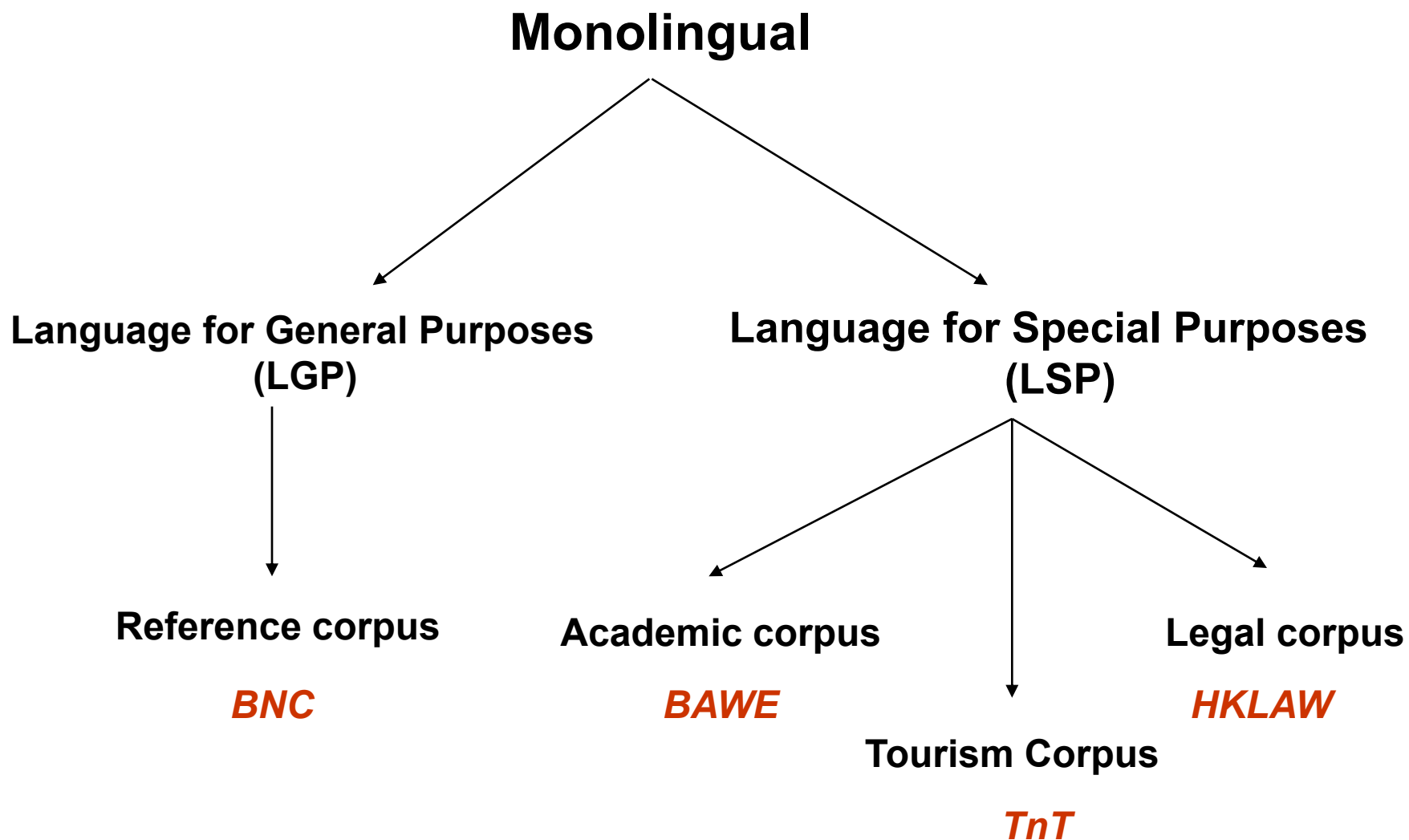
Written

Monolingual

Bi-/Multi-lingual

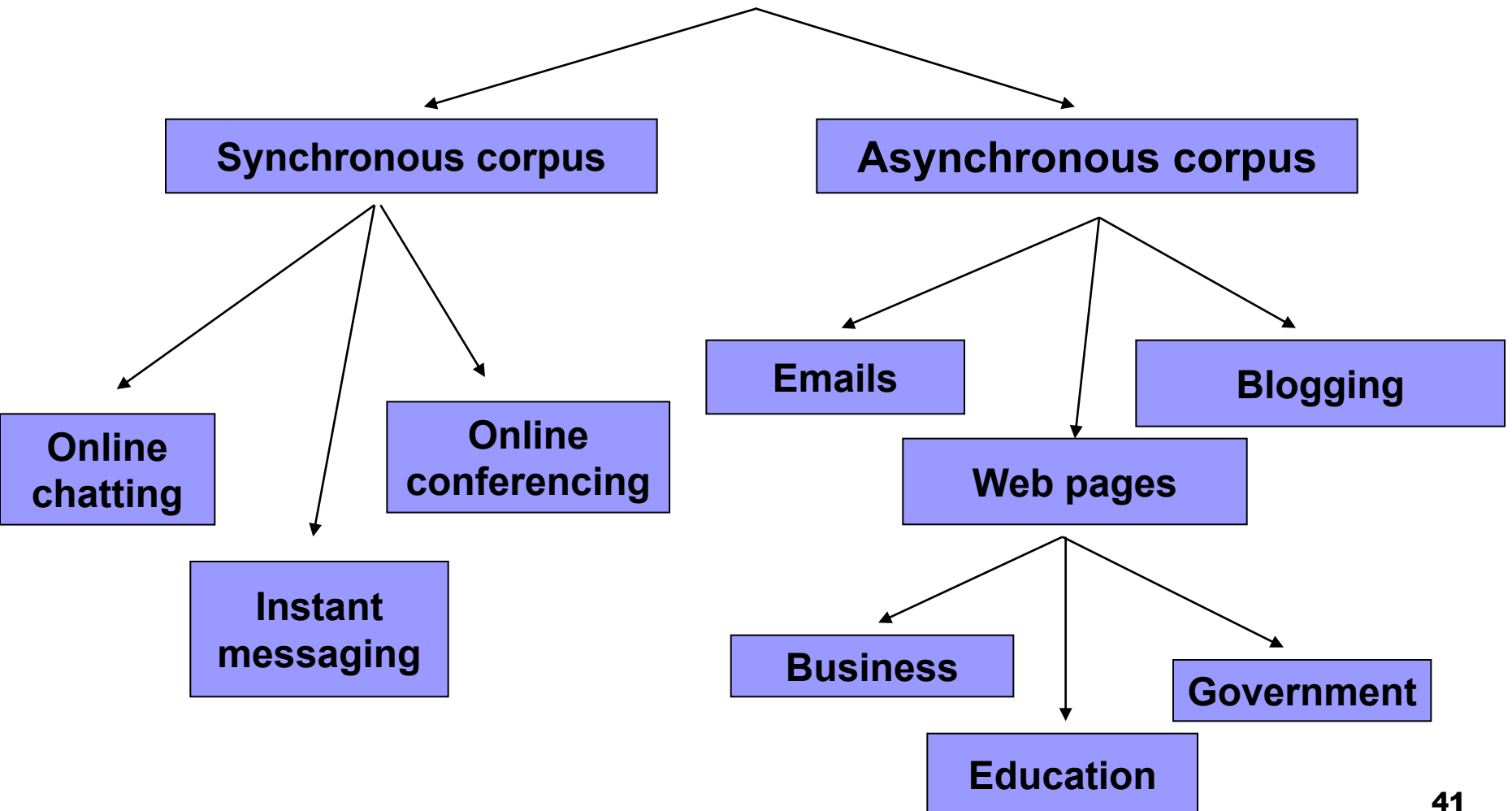
Asynchronous: Emailing; blogging; BBS forum
posting; Online film/book reviewing; etc

Classification of Corpora (Content)



Classification of Corpora (CMC)

CMC Corpora





Methods of corpus-based analysis

- Wordlists
- Concordances
- Collocations
- Keywords



Analytical procedure: Four steps

Step 1: Word listing and counting – Tearing the text apart

Step 2: Compiling a concordance – Putting words back into context

Step 3: Sorting the context in a concordance – Uncovering patterns

Step 5: Examining the context of a word – Looking for collocations




A problem for you

What verbs go with “battle”?

What adjectives go with “battle”?

What phrases contain “battle”?

- 
- “The ability to examine large text corpora in a systematic manner allows access to a quality of evidence that has not been available before” (Sinclair 1991: 4)

‘battle’: LTP Dictionary of Selected Collocations

- **Verbs to the left:** *engage in, fight, force, go into, join in, lose, take part in, win ~*
- **Verbs to the right:** *~ continues, dragged on, ended in stalemate, is in progress, raged*
- **Adj:** *bitter, bloody, crucial, decisive, fierce, final, hopeless, important, last-ditch, long, long-running, major, mock, pitched, real, relentless, running, successful ~*
- **Phrases:** *fight a losing ~, outcome of ~*

BNC Written

- In 90 million words, “battle” comes over 6,000 times, once every 14,000 words.
- Collocated verbs in top 100 linked by MI score:
 - fought (153)/fighting (93)
 - rages (5)/raged (12)
 - waged (10)/waging (12)
 - ensued (8)/ensuing (13)
 - defeated (39)
 - losing (68)
 - won (152)
 - commence (5)

Clusters in BNC Written

- to do **battle** (54)
- fighting a losing **battle** (24)
- win the **battle** (22)
- won the **battle** (22)
- fighting a losing **battle** (21)
- to fight a **battle** (15)

Lexicography and corpora

- Corpus provides **authentic uses** of language
- Extract samples (concordance) to identify different senses
- Word **frequency** information
- Help identify **collocation, set phrase**
Set phrase: *night and day, black and white*
- Modern English dictionaries are all now corpus-based.
Oxford, Collins, Longman, Cambridge...

Linguistics and Corpora

- **Verify** linguistic theory and hypotheses
- Research on empirical linguistics
- Language variation

e.g **Intonation**

- Cheng et al (2008). *A corpus-driven study of discourse intonation*

Grammar

- Biber et al (1999) *Longman Grammar of Spoken and Written English*
- Hunston et al (1996) *Pattern Grammar*

Discourse

- Baker (2006). *Using corpora in discourse analysis.*

Language variation

- Reppen et al (Eds.) (2002). *Using corpora to explore linguistic variation*

Language Teaching and Corpora

- Use corpus as a resource Knowledge :
 - Know better about English:
e.g. answer specific questions of certain words, phrases, structures.
 - Know where the problems are:
e.g. error analysis on a learner corpus
 - Know what should be taught
e.g. syllabus design, teaching materials

Language Teaching and Corpora

■ Use corpus for 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?

Corpus and literary study

- Corpus of literary works:
 - e.g. Corpus of Shakespeare's drama
- Stylistic studies
 - Compare the works of different writers
 - Compare the literary works of different genres, for different readership
- Historical studies
 - Compare works of different historical period
 - Investigate the changes of the patterns of language uses
 - Examine the changes of vocabulary

Corpus and Translation Study

- Corpora as a resource for translation
- Parallel corpora: corpus of translation and its original texts
 - Provide examples of translation
- Corpus of translation vs Corpus of target language
 - Help editing translation to be native-like
 - Help understanding difficult words/concepts