Unit7 Using available corpora

7017115 新居小春

Unit7 Using available corpora

7.1 Introduction
7.2 General Corpora
7.3 Specialized Corpora
7.4 Written Corpora
7.5 Spoken Corpora
7.6 Synchronic Corpora
7.7 Diachronic Corpora
7.8 Learner Corpora
7.9 Monitor Corpora
7.10 Unit Summary and Looking Ahead

7.1 Introduction

- Introduce some of **major publicity available corpus resources**
- Explore **the pros and cons of using ready-made corpora**
- The usefulness of a ready-made corpus must be judged with regard to **the purpose** to which a user intends to put it
- Many of corpora are created for **specific research projects** (=**not publicly available**)

7.1 Introduction

- Focus on major English corpora
- Classified in terms of their **potential use**(includes some **overlap in classification**);

➢**General**(7.2) vs. **Specialized**(7.3) corpus

➢**Written**(7.4) vs. **Spoken**(7.5) corpus

➢**Synchronic**(7.6) vs. **Diachronic**(7.7) corpus

➢**Learner** corpus(7.8)

➢**Monitor** corpus(7.9)

7.2 General Corpora

- **Balanced** with regard to **the variety of a given language**
- Term ‘**balance**’

➢Relative and closely related to a particular research question, if the corpus question claims to be general in nature

➢will be balanced with regard to **genres and domains**

➢**represent** the language under construction

- **Contain written data, spoken data or both**

7.2 General Corpora

- **British National Corpus(BNC)**(1)
  - Comprises **100,106,008 words**
  - Organized in 4,124 written texts and transcripts of speech
  - Designed to represent as wide range of modern British English as possible
  - Written section(90%): includes many others kinds of text
  - Spoken section(10%): 863transcripts of a large amount of informal conversation

7.2 General Corpora

- **British National Corpus(BNC)**(2)

➢POS information

➢Annotated with rich metadata encoded according to the TEI guidelines, using ISO standard 8879(SGML)

➢Internationally agreed standards for encoding

➢A useful resource for a very wide variety of research purpose

➢Access BNC; online(using SARA client program), BNC Web interface, (with a local copy of the corpus,)stand-alone exploration tools(e.g. WordSmith Tools)

7.2 General Corpora

- **British National Corpus(BNC)**(3)

➢In combination with corpora of other languages adopting a similar sampling frame, BNC can provide a reliable basis for contrastive language study

- Designed as matches for the BNC

➢**American National Corpus(ANC)**

➢**Korean National Corpus**

➢**Polish National Corpus(PNC)**

7.3 Specialized Corpora

- **Specialized relative to a general corpus**
- Can be **domain or genre specific** and designed to represent a sublanguage
- Specialized corpora can be **extracted from general corpora**
- E.g. domain specific corpus

➢**Guangzhou Petroleum English Corpus**;

contains 411,612 words of written English from petrochemical domain

➢**HKUST Computer Science Corpus**;

1 million- word corpus of written English sampled from undergraduate textbooks in computer science

7.3 Specialized Corpora

- Recently much interest in the creation and exploitation of specialized corpora in academic or professional settings

- **Corpus of Professional Spoken American English (CPSA)**;

➢Constructed from a selection of transcripts of interactions in professional settings

➢Contains 2 main subcorpora of 1 million words each;

① Consists mainly of academic discussions

② Contains transcripts of White House press conferences

7.3 Specialized Corpora

- **Michigan Corpus of Academic Spoken English（MICASE）**;

➢ Approximately 1.7 million words (nearly 200h of recordings)

➢ Focusing on contemporary university speech within the domain of the University of Michigan

➢ Entire corpus can be accessed online at the corpus website

- **Professional English Research Consortium (PERC)**;

➢ Aims to create a 100-milliom-word discourse used by working professionals and professionals-in-training

➢ Covering a wide range of domain

7.3 Specialized Corpora

- Language may vary considerably across domain and genre

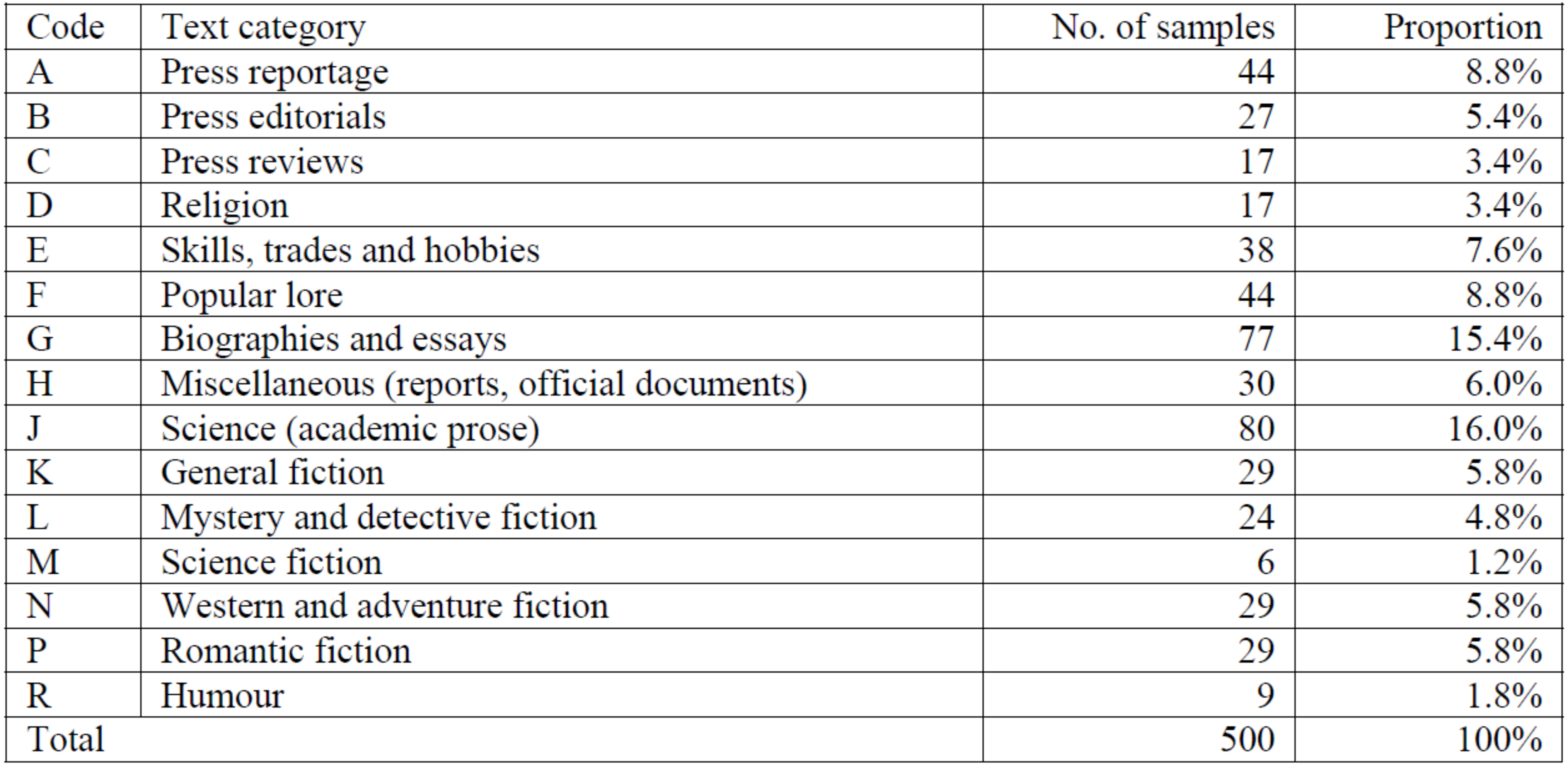
➢Specialized corpora provide **valuable resources for investigations in the relevant domains and genres**

7.4 Written Corpora

- **Brown University Standard Corpus of Present-day American English (Brown corpus)**;
  - 1st modern corpus of English
  - Corpus of written American English
  - Complied using 500 chunks of **approximately 2,000 words of written texts**
  - Sampled from **15 categories**

7.4 Written Corpora

Table 7.1 Text categories in the Brown corpus



Code	Text category	No. of samples	Proportion
A	Press reportage	44	8.8%
B	Press editorials	27	5.4%
C	Press reviews	17	3.4%
D	Religion	17	3.4%
E	Skills, trades and hobbies	38	7.6%
F	Popular lore	44	8.8%
G	Biographies and essays	77	15.4%
H	Miscellaneous (reports, official documents)	30	6.0%
J	Science (academic prose)	80	16.0%
K	General fiction	29	5.8%
L	Mystery and detective fiction	24	4.8%
M	Science fiction	6	1.2%
N	Western and adventure fiction	29	5.8%
P	Romantic fiction	29	5.8%
R	Humour	9	1.8%
Total		500	100%

7.4 Written Corpora

- A number of corpora which follow the Brown model;

- **Lancaster-Oslo-Bergen Corpus of British English**(**LOB**)
  - **British match** for the Brown corpus
  - Created using **the same sampling techniques** with the exception (LOB aims to represent written British English used in 1961)

7.4 Written Corpora

- Brown and LOB

➢Provide an **ideal basis for the comparison of the 2 major varieties of English as used in the early 1960s**

➢POS tagged

➢Sub-samples have been parsed

7.4 Written Corpora

- 2 Freiburg corpora

① **Freiburg-LOB Corpus of British English(FLOB)**

② **Freiburg-Brown Corpus of American English(Frown)**

➢Available to **mirror the Brown/LOB relationship in the early 1990s rather than 1960s**

➢Represent written British and American English as used in 1991

➢Enable users to **track language changes** in British and American English over the intervening 3decades between Brown/LOB and FLOB/Frown

7.4 Written Corpora

- Corpora for varieties of English using the Brown sampling model;
- **Australian Corpus of English (ACE)**

➢written Australian English from 1986 and after

- **Wellington Corpus (WWC)**

➢written New Zealand English, covering the years between 1986 and 1990

- **Kolhapur Corpus**

➢Indian English dating from 1978

7.4 Written Corpora

- **Lancaster Corpus of Mandarin Chinese (LCMC)**

➢Sampling frame cross languages/language varieties

➢Chinese match for the FLOB and Frown corpora

➢Makes it possible to contrast Chinese with 2 major English varieties

7.5 Spoken Corpora

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

➢Corpus of **spoken British English** dating from the 1960s to the mid-1970s

➢Consists of 100 texts, each of 5,000 words, totaling half a million running words

➢Distinction is made between **dialogue** and **monologue** in the organization of the corpus

➢Prosodically annotated

7.5 Spoken Corpora

- **Lancaster/IBM Spoken English Corpus (SEC)**

➢Consists of approximately 53,000 words of **spoken British English**

➢Mainly taken from radio broadcasts dating from the mid-1980s and covering a range of speech categories

➢Available in an orthographically transcribed form and POS tagged, parsed or prosodically annotated version

7.5 Spoken Corpora

- **Cambridge and Nottingham Corpus of Discourse in English (CANCODE)**

➢Part of the Cambridge International corpus (CIC)

➢Comprises 5 million words of **transcribed spontaneous speech collected in Britain** between 1995 and 2000

➢Covering a wide variety of situations

➢**Coded with information pertaining to the relationship between the speakers**, which allows users to look more closely at **how different levels of familiarity (formality) affect the way in which people speak to each other**

7.5 Spoken Corpora

- **Santa Barbara Corpus of Spoken American English (SBCSAE)**

➢Part of the USA component of the International Corpus of English (ICE)

➢Based on hundreds of **recordings of spontaneous speech** from all over the US

➢Reflects the many ways that people use language in their lives

➢Particularly useful for research into **speech recognition** as each speech file is accompanied by a transcript in which phrases are time stamped to allow them to be linked with the audio recording from which the transcription was produced

7.5 Spoken Corpora

- **Wellington Corpus of Spoken New Zealand English (WSC)**

➢Comprises 1 million words of **spoken New Zealand English** in the from of 551 extracts collected between 1988 and 1944

➢Formal speech/monologue accounts for 12% of the data, semi-formal speech/elicited monologue 13% while informal speech /dialogue accounts for 75%

➢**The unusually high proportion of private material makes the corpus a valuable resource for research into informal spoken registers**

7.6 Synchronic Corpora

- Compare various English

➢E.g. **Brown family**

➢Results from comparison must be interpreted with caution where the corpora under examination were built to represent English in different time periods or the Brown model has been modified

7.6 Synchronic Corpora

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

➢Specifically designed for the **synchronic study of world Englishes**

➢Consists of a collection of 20 corpora of 1 million words each, each composed of written and spoken English produced after 1989 in countries or regions in which English is a 1st or major language

➢Primary aim: to facilitate comparable studies of English used worldwide(each component follows a common corpus design as well as a common scheme for grammatical annotation to ensure comparability among the component corpora)

➢Encoded at various levels, including textual markup, POS tagging and syntactic parsing

7.6 Synchronic Corpora

- Considerably fewer corpora available for regional dialects than national varieties

➢Comparing dialects is assumed to be less meaningful

- **Spoken component if the BNC**

➢Allow users to compare dialects in Britain

- **Longman Spoken American Corpus**

➢Built to match the demographically sampled component of the spoken BNC

➢Can be used to compare regional dialects in the USA

7.6 Synchronic Corpora

- Spoken corpus of the **Survey of English dialects (SED)**

➢Specifically for the study of **English dialects**

➢Started in 1948

➢Initially comprised a questionnaire-based survey of traditional dialects based on extensive interviews from 318 locations all over rural England

➢Recordings made during 1948-1973 consist of about 60h of dialogue of elderly people, transcribed with sound files to transcripts

➢Marked up in TEL-compliant SGML and POS tagged using CLAWS

7.6 Synchronic Corpora

- Presently few **synchronic corpora suitable for studies of dialects and varieties for languages other than English**

- **Linguistic Variation in Chinese Speech Communities (LIVAC)**

➢Contains **texts from representative Chinese newspapers and the electronic media of 6 Chinese speech communities**

➢Collection of materials from these diverse communities is synchronized so that all of the components are comparable

➢Under construction (some samples are already available)

7.7 Diachronic Corpora

- Contains texts from the same language gathered from different time periods (far more extensive)
- Used to track changes in language evolution
- Typically contains only written language for practical reasons

- Corpora of speech from earlier periods;

➢ **Helsinki Dialect Corpus**

➢ **Corpus of English Dialogues 1560-1760**

7.7 Diachronic Corpora

- **Helsinki Diachronic of English Texts (the Helsinki Corpus)**

- Consists of approximately 1.5 million words of English in the from of 400 text samples, dating from the 8th to 15th centuries
- Covers a wide range of genres and sociolinguistic variables
- Divided into 3 periods (Old, Middle, and Early Modern English) and 11subperiods

7.7 Diachronic Corpora

- **A Representative Corpus of Historical English Registers corpus (ARCHER corpus)**

➢Covers both British and American English dating from 1650 to 1990, dividing into 50-year periods

➢Including spoken data from the later period

- **Lampeter Corpus of Early Modern English Tracts**

➢Contains 1.1 million words of pamphlet literature dating from 1640 and 1740

➢Includes whole texts (useful for the study of textual organization in Early Modern English)

7.8 Learner Corpora

- A collection of the writing or speech of **learners acquiring a second language(L2)** (‘developmental data’)
- Used for either cross-sectional or longitudinal analysis
- Opposed to a ‘**developmental corpus**’

➢Consists of data **produced by children acquiring their first language(L1)**

➢Specifically for L1 data as opposed to learner corpus

➢**Child Language Data Exchange System (CHILDES)**

➢**Polytechnic of Wales Corpus (POW)**

7.8 Learner Corpora

- **International Corpus of Learner English (ICLE)**

➢Contains approximately 3 million words of essays written by advanced learners of English from 14 different mother tongue backgrounds

➢Error and POS tagged version available in near future

➢Used in combination with the Louvain Corpus of Native English Essays (LOCNESS) to **compare native and learner English**

➢Available for linguistic research but cannot be used for commercial purposes

7.8 Learner Corpora

- **Longman Learners' Corpus**

➢Contains 10 million words of text written by students of English at a range of levels of proficiency from 20 different L1 backgrounds

➢Elicitation task (used to gather the texts varied)

➢Each essay is coded by L1 background and proficiency level, amongst other things

➢Not tagged for POS, but part of the corpus are manually error-tagged (only for internal use at Longman Dictionaries)

➢Invaluable information about the mistakes students make and what they already know

➢Publicly available for commercial purposes

7.8 Learner Corpora

- Cambridge Learner Corpus (CLC)

➢A large collection of **examples of English writing from learners of English all over the world**

➢Contains over 20 million words and expanding continually

➢Comes from analyzed exam scripts written by students taking Cambridge ESOL English exams worldwide

➢Contains 50,000 scripts from 150 countries (100 different L1)

➢Coded with information about students

➢Not publicly available

7.8 Learner Corpora

- Corpora covering only one L1 background
- **HKUST Corpus of Learner English**
  - 10-million-word corpus composed of **written essays and exam scripts** of **Chinese leaners in Hong Kong**
- **Chinese Learner English Corpus (CLEC)**
  - Contains 1 million words from **writing** produced by 5 types of **Chinese learners** ranging from middle school students to senior English majors
- **NICT JLE Corpus**
  - Contains 1 million words of error tagged **spoken English** produced by **Japanese learners**

7.8 Learner Corpora

- **Japanese EFL Learner (JEFLL) Corpus**
  - 1-million-word corpus containing 10,000 **sample essays** written by **Japanese learners** from Years 7-12 in secondary schools
- **Janus Pannonius University(JPU) Learner Corpus**
  - 400,000 words, containing the **essays** of advanced level **Hungarian university students** that were collected from 1992 to 1998
- **Uppsala Student English (USE) Corpus**
  - Contains 1 million words of **texts** written primarily by **Swedish university students** who are advanced learners of English with a high level of proficiency

7.8 Learner Corpora

- **Polish Learner English Corpus**

➢Designed as a half-a-million-word corpus of **written** learner data produced by **Polish learners** from a range of learner styles at different proficiency levels, from begging to post-advanced

7.9 Monitor Corpora

- **Sample corpora**: introduced in the previous sections are constant in size
- **Monitor corpus**: constantly supplemented with fresh material and keeps increasing in size (the proportion of text types included remains constant), typically much larger then sample corpora
- **Bank of English (BoE)**

➢Widely acknowledged to be an example of a monitor corpus

➢Increased in size progressively since its inception in the 1980s

➢Around 524 million words at present

7.9 Monitor Corpora

- **Global English Monitor Corpus**

➢Started in late 2001 as an electronic archive of the world's leading newspapers in English

➢Expected to reach billions of words within a few years

➢Aims at monitoring language use and sematic change in English as reflected in newspapers so as to allow for research into whether the English language discourses in Britain, the US, Australia, Pakistan and South Africa are convergent or divergent over time

7.9 Monitor Corpora

- **Analysis of Verbal Interaction and Automated Text Retrieval (AVIATOR)**

➢Automatically monitors language change, using a series of filters to identify and categorize new word forms, new word pairs or terms, and change in meaning

- Monitor corpus **does not have a finite size** (‘ongoing archive’ rather than a true corpus)

7.9 Monitor Corpora

- Impromptu debate at a joint conference of the Association for the Literary and Linguistic Computing (ALLC) and the Association for Computing in the Humanities (AHC)

- ➢Ouirk and Leech: arguing in favor of **the balanced sample corpus model**
- ➢Sinclair and Meijis: arguing in favor of **the monitor corpus model**
- ➢Monitor corpus team won the debate in 1992
- ➢**Sample corpus has won the wider debate, as evidenced by it being the dominant tradition in modern corpus building**
- ➢Majority of corpora are **balanced sample corpora**, as exemplified by the pioneer English
- ➢Idea of the monitor corpus is still important and deserves a review

7.9 Monitor Corpora

- First developed in Sinclair
- ➢Argued against static sample corpora and in favor of an **ever growing dynamic monitor corpus**
- ➢Amend his views ‘**as new advances come on stream**’ and no longer holds many of the positions (expressed still have some currency)
- ➢Worth reviewing the arguments presented by Sinclar against sampled corpora
- ➢Major arguments relate to the overall **corpus size** and the **sample size** (have largely been neutralized by an increase in both computer power and the availability of electronic texts)
- ➢‘the continued growth in the size of corpora has generally implied an increase in sample sizes as well as the number of samples’

7.9 Monitor Corpora

- even in corpora consisting of 2,000-word samples, **frequent linguistic features are quite stable both within samples and across samples in different text categories**
- For relatively rare features and for vocabulary, though, Sinclair's warning is still valid
- Monitor corpus undergoes a **partial self-recycling after reaching some sort of saturation**
- the inflow of the new data is subjected to an **automatic monitoring process** which will only admit new material to the corpus where that material shows some features which differ significantly from the stable part of the corpus

7.9 Monitor Corpora

- difficulties with the monitor corpus model(1)

➢as this approach **rejects any principled attempt to balance a corpus, depending instead upon sheer size to deal with the issue** (monitor corpora are a less reliable basis than balanced sample corpora for quantitative (as opposed to qualitative) analysis)

➢as this approach argues strongly **in favor of whole texts, text availability becomes a difficulty in the already sensitive context of copyright**

7.9 Monitor Corpora

- difficulties with the monitor corpus model(2)

➢**quite confusing to indicate a specific corpus version with its word count** (under such circumstances it is only the corpus builders, not the users, who know what is contained in a specific version)

➢as a monitor corpus keeps **growing in size, results cannot easily be compared** (thus loses its value as a standard reference)

7.9 Monitor Corpora

- a **dynamic monitor corpus** should in effect **consist of a series of static corpora over time**
- One final concern of the dynamic model

➢even if the **huge corpus size and required processing speed** should not become a problem as the rapid development of computing technology makes this a non-issue

➢there is no guarantee that the same criteria will be used to filter in new data in the long term, meaning that even if a diachronic archive of the sort suggested is established, the comparability of the archived version of the corpus would be in doubt.

7.9 Monitor Corpora

- **Primarily designed to track changes from different periods**
- similar to a diachronic corpus
- normally covers a **shorter span of time** but in a **much more fine-grained fashion** than the diachronic corpora discussed so far.
- It is possible, using a monitor corpus

➢to **study relatively rapid language change**

➢to **study change happening at a much slower rate**

7.10 Unit Summary and Looking Ahead

- the usefulness of a given corpus **depends upon the research question a user intends to investigate using that corpus**
- researchers and students will find that they are **not able to address their research questions using ready-made corpora**
- one must **build one's own corpus**
- In the unit that follows, we will discuss the **principal considerations involved in building such DIY corpora**