Chapter 2
Language Resources

As for corpus-based studies, the research questions, observations and findings are largely dependent on the language resources. General corpora usually provide a diverse range of genres and registers, whereas specialised corpora have an exclusive focus. It is also worth noting that language resources are not restricted to different types of corpora, either as the target corpus or as the reference corpus. Very often, corpus-based investigations would turn to lexical resources for further analysis of the data obtained from the chosen corpus or corpora. In this chapter we are going to briefly review some of the well-known language resources, among which some are used in our studies (e.g. BNC, ICE) while some (e.g. PubMed, WordNet) are commonly used in the computation of language features. We should admit that it would be difficult to provide a comphrehensive list of all the important resources or to mention all the studies that have been done on the basis of the resources. Therefore, the overview of language resources in this chapter will be focused on the design of the selected resources, their intended purposes and major application.

2.1 General Corpora

2.1.1 The Brown Corpus and the Brown Family

The Brown Corpus It would be safe to say that no introduction to corpora would fail to mention the Brown Corpus, the first publicly available computerised, general corpus. Better known as the Brown Corpus, the Standard Corpus of Present-Day Edited American English was compiled by W. Nelson Francis and Henry Kučera and first released in 1964. The manual available at http://icame.uib.no/brown/bcm.html reveals the following features of the corpus:

1. The corpus consists of edited English prose printed in the USA during 1961.
2. A rough count of 2000 words was made for each sample.
3. Five hundred samples were chosen for their representative quality.

A. C. Fang, J. Cao, *Text Genres and Registers: The Computation of Linguistic Features,*
DOI 10.1007/978-3-662-45100-7_2

4. The samples represent 15 different categories of prose.
5. The corpus will be used for comparative studies.

See Table 2.1 for the composition of the Brown Corpus.

The Brown Family The Brown Corpus is well recognised not only because it is the first computerised general corpus but also because the structure of the Brown composition has been cloned in a set of corpora that are often referred to as the 'Brown Family', which includes the four core members (i.e. Brown, LOB, Frown and FLOB) and the extended family members as well. See Table 2.2 for a summary, where the corpora are arranged according to the date of their first release. The members of the Brown family not only copied the composition of the Brown Corpus and the size of the sample text (about 2000 words each), and they are also samples of printed materials, or written English.

Application

a. Linguistic Studies

As McEnery et al. (2006) pointed out, 'lexical and grammatical studies are probably the areas that have benefited most from corpus data' (p. 145). With the availability of the Brown family, substantial studies have been made on various aspects both intra-corpus and inter-corpora.

In terms of intra-corpus studies[1], the most typical study would be frequency investigation (e.g. Zettersten and Kučera 1978; Francis and Kučera 1982; Nakamura 1989, 2002), and substantial studies have also been made in linguistics, including lexical and grammatical studies (e.g. Johansson 1978; Ellegård 1978; Kjellmer 1979, 1980), semantic studies (e.g. Hermerén 1978; Warren 1978), and also studies on collocations (e.g. Backlund 1981; Kjellmer 1982).

Inter-corpora studies have been firstly focus on the comparison between American and British English since for a long time the Brown and LOB corpora have been the only available comparable language resources. Again, the studies consist of frequency investigation (e.g. Johansson 1980; Hofland and Johansson 1982; Krogvig and Johansson 1984), lexical and grammatical analysis (e.g. Krogvig and Johansson 1981; Johansson and Norheim 1988; Collins 1996), and syntax and semantic analysis (e.g. Coates and Leech 1980; Coates 1983; Johansson and Oksefjell 1996). With the availability of the Brown family, comparative studies start to cover more variations of English, for instance, a comparison between American, British and Indian English (e.g. Leitner 1994).

In addition to the aforementioned synchronic studies, diachronic comparison has also been made across the four core members of the Brown family (i.e. Brown, LOB, Frown, and FLOB), such as historical syntactic investigation in general (Rissanen 2012), and more specifically on English adverbial subordinators (Rissanen 2011).

[1] http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/bibliography.html.

Table 2.1 Composition of the Brown Corpus

Writing (100%)				
Informative prose			Imaginative prose	
A: Press: Reportage	Political	14	*K: General fiction*	
	Sports	7	Novels	20
	Society	3	Short stories	9
	Spot news	9		
	Financial	4	*L: Mystery and detective fiction*	
	Cultural	7	Novels	20
B: Press: Editorial	Institutional	10	Short stories	4
	Personal	10		
	Letters to the editor	7	*M: Science fiction*	
C: Press: Reviews		17	Novels	3
D: Religion	Books	7	Short stories	3
	Periodicals	6		
	Tracts	4	*N: Adventure and western fiction*	
E: Skills and hobbies	Books	2	Novels	15
	Periodicals	34	Short stories	14
F: Popular lore	Books	23		
	Periodicals	25	*P: Romance and love story*	
G: Belles-lettres, etc.	Books	38	Novels	14
	Periodicals	37	Short stories	15
H: Miscellaneous	Government documents	24		
	Foundation reports	2	*R: Humour*	
	Industry reports	2	Novels	3
	College catalogue	1	Essays, etc.	6
	Industry house organ	1		
J: Learned	Natural sciences	12		
	Medicine	5		
	Mathematics	4		
	Social and behav-ioural sciences	14		
	Polit, law, education	15		
	Humanities	18		
	Technology and engineering	12		

Table 2.2 Corpora of the Brown family

Corpus	First release	Language	Data period
Brown Corpus	1964	American English	1961
Lancaster-Oslo-Bergen (LOB)[a]	1976	British English	1961
The Kolhapur Corpus of Indian English[b]	1986	Indian English	1978
The Australian Corpus of English (ACE)[c]	1987	Australian English	1986
The Wellington Corpus of Written New Zealand English (WWC)[d]	1993	New Zealand English	1986–1990
Freiburg Update of the Brown Corpus (Frown)[e]	1999	American English	1992
Freiburg-LOB Corpus of British English (FLOB)[f]	1999	British English	1991
CROWN[g]	2012	American English	2009
CLOB[g]	2012	British English	2009/± 1 year

[a] Manual: http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM
[b] Manual: http://khnt.hit.uib.no/icame/manuals/kolhapur/index.htm
[c] Manual: http://khnt.hit.uib.no/icame/manuals/ace/INDEX.HTM
[d] Manual: http://icame.uib.no/wellman/well.htm
[e] Manual: http://khnt.aksis.uib.no/icame/manuals/frown/INDEX.HTM#pre
[f] Manual: http://khnt.aksis.uib.no/icame/manuals/flob/INDEX.HTM
[g] http://www.fleric.org.cn/crown/

b. Dictionary Compilation

The Brown Corpus has also been used as the resource for the compilation of dictionaries. *American Heritage Dictionary* (1969), employing linguistic information (e.g. frequency counts) from the Brown Corpus, can be considered as the first corpus-based dictionary during the computerised corpus era. Later in 1994, *A Dictionary of English Collocations: based on the Brown Corpus* (1994), written by Kjellmer, was published by the Clarendon Press.

2.1.2 The International Corpus of English (ICE) Family

The International Corpus of English (ICE), first proposed by Sidney Greenbaum in 1988, was designed to serve as the language resource for comparative studies of English worldwide. Here *English* refers to the English language used in 24 nations or regions, where it is the first language or an official additional language. Due to the main goal of such a corpus, a general design[2] is expected to be followed by all the corpora in the ICE family:

[2] http://ice-corpora.net/ice/manuals.htm.

1. The overall size of each corpus is one million words of English produced after 1989.
2. Each corpus consists of 500 texts of about 2000 words each.
3. Each corpus covers 300 spoken and 200 written English texts.

See Table 2.3 for the general structure[3] (table quoted from Fang 2007, p. 29).

The ICE launched in 1990, and so far the following 12 subsets of the ICE family are available commercially or free under license:

1. Canada (ICE-CAN)
2. East Africa (ICE-EA)
3. Great Britain (ICE-GB)
4. Hong Kong (ICE-HK)
5. India (ICE-IND)
6. Ireland (ICE-IRE)
7. Jamaica (ICE-JA)
8. New Zealand (ICE-NZ)
9. Singapore (ICE-SIN)
10. Sri Lanka (ICE-SL)
11. The Philippines (ICE-PHI)
12. The USA (written) (ICE-USA)

ICAME Journal No 34 (2010) discusses the creation of new members of the ICE family, including Fiji, Bahamas, Malta and Nigeria. In addition, according to the ICE website (February 2013) , '[t]he tagging of all currently available ICE corpora with CLAWS7 and the USAS semantic tagger is now complete'. Among them, the ICE-GB is tagged and parsed, and manually validated. Tagging and parsing will be discussed later in Chap. 3.

Application

a. Linguistic Studies

The primary goal of ICE project is to facilitate the intercorpus studies between different varieties of Englishes. This section will mainly introduce special volumes devoted to the ICE project.

In 2004, a special issue of *World Englishes* reported the first series of ICE-based studies. The comparisons are made between inner circle varieties (e.g. British or New Zealand English) and outside circle varieties (e.g. Hong Kong, Indian and Singapore English). Linguistic features include multi-word verbs (Schneider 2004), negation of lexical *have* (Nelson 2004) and article use (Sand 2004).

A most recent book *Mapping Unity and Diversity Worldwide: Corpus-based Studies of New Englishes* edited by Hundt and Gut (2012)[4] can be considered a second series of the ICE-based studies. Again, varieties from the outside circle are compared with those from the inner circle, and language use has been examined from various

[3] http://ice-corpora.net/ice/design.htm.

[4] http://benjamins.com/#catalog/books/veaw.g43/main.

Table 2.3 Composition of ICE-GB

Speech (60 %)				Writing (40 %)			
Dialogue	*Private*			Nonprinted	*Student writing*		
	S1A1	Direct conversations	90		W1A1	Untimed essays	10
	S1A2	Distanced conversations	10		W1A2	Timed essays	10
	Public				*Correspondence*		
	S1B1	Class lessons	20		W1B1	Social letters	15
	S1B2	Broadcast discussions	20		W1B2	Business letters	15
	S1B3	Broadcast interviews	10	Printed	*Informational*		
	S1B4	Parliamentary debates	10		W2A1	Learned: humanities	10
	S1B5	Legal cross-examinations	10		W2A2	Learned: social sciences	10
	S1B6	Business transactions	10		W2A3	Learned: natural sciences	10
Monologue	*Unscripted*				W2A4	Learned: technology	10
	S2A1	Spontaneous commentaries	20		W2B1	Popular: humanities	10
	S2A2	Unscripted speeches	30		W2B2	Popular: social sciences	10
	S2A3	Demonstrations	10		W2B3	Popular: natural sciences	10
	S2A4	Legal presentations	10		W2B4	Popular: technology	10
	Mixed				W2C1	Press news reports	20
	S2B1	Broadcast news	20		*Instructional*		
	Scripted				W2D1	Administrative writing	10
	S2B2	Broadcast talks	20		W2D2	Skills and hobbies	10
	S2B3	Nonbroadcast talks	10		*Persuasive*		
					W2E1	Press editorials	10
					Creative		
					W2F1	Fiction	20

aspects, including verbs (Nelson and Ren 2012; Schilk et al. 2012; Schneider and Hundt 2012; Zipp and Bernaisch 2012), modals (Auwera et al. 2012; Collins and Yao 2012; Deuber et al. 2012), progressives (Hilert and Krug 2012), relativization strategies (Gut and Coronel 2012), infinitives (Mair and Winkle 2012) and quotatives (Höhn 2012).

In addition, as pointed out at the ICE website[5], 'for most participating countries, the ICE project is stimulating the first systematic investigation of the national variety'. Therefore, empirical studies have also contributed to our understanding of 'New' Englishes, such as Indian (e.g. Lange 2012), African (e.g. Jeffery and Van Rooy 2004; Nelson and Ren 2012) and Asian Englishes (e.g. Auwera et al. 2012).

b. Grammar Book Compilation

Another most important outcome from the ICE family is the production of grammar books, namely,

Oxford English Grammar (Greenbaum 1996)
An Introduction to English Grammar (3rd ed.) (Greenbaum and Nelson 2009)
Oxford Modern English Grammar (Aarts 2011)

2.1.3 BNC and ANC

2.1.3.1 The British National Corpus (BNC)

The British National Corpus (BNC), a financially available corpus, was built on the efforts of an academic-industrial consortium, including Oxford University Press, Longman Group Ltd, Chambers Harrap, Oxford University Computing Services, Lancaster University and British Library Research and Development Department. The BNC is a 100-million-word corpus designed to represent contemporary English, and the main features can be summarised as follows:

1. It consists of both written (90%) and spoken (10%) samples.
2. The texts are mainly from the period of 1985 to 1994.
3. Written texts are selected from newspapers, periodicals, journals, books, student essays, letters and other sources.
4. Spoken texts are collected in different contexts with speakers from a balanced demographic background.

These features have categorised the BNC as a large, balanced, general corpus. Till now, there are three versions available as listed in Table 2.4.

More importantly, in the planning of the compilation of the BNC, quite a few applications have been laid out, as listed in the *BNC User Reference Guide*[6]:

[5] http://ice-corpora.net/ice/index.htm.

[6] http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html.

Table 2.4 BNC versions

Versions	First release	Distribution	Features
BNC 1.0	1995	European researchers	
BNC World	2001	Worldwide	Tagged
BNC XML	2007	Worldwide	Tagged, XML

- Reference book publishing
- Academic linguistic research
- Language teaching
- Artificial intelligence
- Natural language processing
- Speech processing
- Information retrieval

According to the *BNC User Reference Guide*, the BNC contains eight different text categories based on the categorizations by Lee (2001). Table 2.5 summarises the composition of the BNC.

Application

a. Linguistic Studies

The major contribution of the BNC is the application in linguistic studies, including investigations into morphological features (e.g. Plag et al. 1999), syntactic features (e.g. Kerz and Haas 2009; Choi 2012), pragmatic features (e.g. Deutschmann 2003; Jucker et al. 2008; Cheng 2010), register variations (e.g. Takahashi 2006) and also sociolinguistic studies (e.g. Xiao and Tao 2007).

With 100 million words, the BNC has been used to create frequency lists (e.g. Rayson and Garside 2000; Leech et al. 2001;Wang 2005); being a general corpus, the BNC has also been used as a reference corpus (e.g. Louwerse et al. 2008).

b. English Education

The BNC has been used in language teaching in mainly two ways: direct and indirectly. Texts from the BNC have been directly used as the material for English learners with the goal to improve their English proficiency (e.g. Aston 1998; Miangah 2011). Indirectly, the BNC has been used as the native English corpus to evaluate the English of non-native speakers (e.g. Chujo 2004; Liu et al. 2008; Mukundan and Roslim 2009; Philip et al. 2012; Sonbul and Schmitt 2013).

c. Artificial Intelligence and Natural Language Processing

The BNC has been contributed in the field of artificial intelligence by providing extensive data. It has been serving as the test bed for a variety of experiments, such as automatic acquisition of topic signatures (e.g. Cuadros et al. 2006), text semantic similarity (e.g. Mihalcea et al. 2006), nontopical classification of documents (e.g. Bekkerman et al. 2006) and automatic extraction of concept–feature triples (e.g. Kelly et al. 2010).

Table 2.5 Composition of BNC. (Adapted from http://www.natcorp.ox.ac.uk/docs/URG/codes.html#classcodes)

Speech (10%)			Writing (90%)		
Broadcasting	Discussion	53	*Academic prose*	Humanities_arts	87
	Documentary	10		Medicine	24
	News	12		Nat_science	43
				Polit_law_edu	186
Classroom		58		Soc_science	142
				Tech_engin	23
Consultations		128	*Admin*		12
			Advertisements		59
Conversations		153	*Biography*		100
			Commerce		112
Courtroom		13	*Email*		7
			Essays	School	7
Demonstrations		6		University	3
			Fiction	Drama	2
Interviews		13		Poetry	30
				Prose	431
Interview oral history		119	*Hansard*		4
			Institute doc		43
Lectures	Commerce	3	*Instructional*		15
	Humanities_arts	4	*Letters*	Personal	6
	Nat_science	4		Professional	11
	Polit_law_edu	7	*Miscellaneous*		503
	Soc_science	13	*News script*		32
			Newsp_brd-sht_nat	Arts	51
Meeting		132		Commerce	44
				Editorial	12
Parliamentary		6		Miscellaneous	95
				Report	49
Pub_debate		16		Science	29
				Social	36
Sermon		16		Sports	24

Table 2.5 (continued)

Speech (10%)	Writing (90%)		
	Newsp_other	Arts	15
		Commerce	17
		Report	39
		Science	23
		Social	37
		Sports	9
	Newsp_tabloid		6
	Nonacademic	Humanities_arts	110
		Medicine	17
		Nat_science	62
		Polit_law_edu	93
		Soc_science	123
		Tech_engin	123
	Pop_lore		211
	Religion		35

The BNC has also been used in the field of natural language processing for evaluations of models, such as text genre detection (e.g. Stamatatos et al. 2000), automatic term extraction (e.g. Kit and Liu 2008), *n*-grams for search engine (e.g. Keller and Lapata 2003), semantic graph (e.g. Widdows et al. 2002), speech recognition (e.g. Goyoh and Renals 1999; Athanaselis et al. 2005).

2.1.3.2 The American National Corpus (ANC)[7]

The ANC project began in 1988 for the goal of creating an American counterpart of the BNC, at least 100 million words of contemporary American English represented in a spectrum of genres. Similar to the BNC, the ANC consists of both spoken and written texts. The unique features of the ANC are as follows:

1. Sample texts are from the year of 1990 onward.
2. Each sample text is at least 1000 words.
3. All data are marked up with multi-layer annotations, including structural mark-up, sentence boundaries, part-of-speech (POS) tags, noun chunks, verb chunks and named entities.
4. All the data and annotations are free.

The ANC was designed to serve for the purposes of education, linguistic research and technology development.

[7] http://www.americannationalcorpus.org/index.html.

http://www.springer.com/978-3-662-45099-4

Text Genres and Registers: The Computation of Linguistic Features

Fang, C.A.; Cao, J.

2015, XIII, 267 p. 40 illus., Hardcover
ISBN: 978-3-662-45099-4