

Contents

I: Corpus linguistics and the BNC	1
1 Corpus linguistics	4
1.1 What is a corpus?	4
1.2 What can you get out of a corpus?	5
1.3 How have corpora been used?	10
1.3.1 What kinds of corpora exist?	10
1.3.2 Some application areas	12
1.3.3 Collocation	13
1.3.4 Contrastive studies	15
1.3.5 NLP applications	18
1.3.6 Language teaching	19
1.4 How should a corpus be constructed?	21
1.4.1 Corpus design	21
1.4.2 Encoding, annotation, and transcription	24
2 The British National Corpus	28
2.1 How the BNC was constructed	28
2.1.1 Corpus design	28
2.1.2 Encoding, annotation, and transcription	33
2.2 Using the BNC: some caveats	36
2.2.1 Source materials	37
2.2.2 Sampling, encoding, and tagging errors	38
2.2.3 What is a BNC document?	39
2.2.4 Miscellaneous problems	40
3 Future corpora	42
II: Exploring the BNC with SARA	45
1 Old words and new words	48
1.1 The problem: finding evidence of language change	48
1.1.1 Neologism and disuse	48
1.1.2 Highlighted features	48
1.1.3 Before you start	49
1.2 Procedure	49
1.2.1 Starting SARA	49
1.2.2 Logging on	49
1.2.3 Getting help	50
1.2.4 Quitting SARA	50
1.2.5 Using Phrase Query to find a word: ‘cracks- men’	51

1.2.6	Viewing the solutions: display modes	51
1.2.7	Obtaining contextual information: the Source and Browse options	52
1.2.8	Changing the defaults: the User preferences dialogue box	54
1.2.9	Comparing queries: 'cracksmen' and 'cracksman'	57
1.2.10	Viewing multiple solutions: 'whammy'	58
1.3	Discussion and suggestions for further work	60
1.3.1	Caveats	60
1.3.2	Some similar problems	61
2	What is more than one corpus?	63
2.1	The problem: relative frequencies	63
2.1.1	'Corpus' in dictionaries	63
2.1.2	Highlighted features	63
2.1.3	Before you start	64
2.2	Procedure	64
2.2.1	Finding word frequencies using Word Query	64
2.2.2	Looking for more than one word form using Word Query	65
2.2.3	Defining download criteria	66
2.2.4	Thinning downloaded solutions	68
2.2.5	Saving and re-opening queries	70
2.3	Discussion and suggestions for further work	72
2.3.1	Phrase Query or Word Query?	72
2.3.2	Some similar problems	72
3	When is ajar not a door?	74
3.1	The problem: words and their company	74
3.1.1	Collocation	74
3.1.2	Highlighted features	75
3.1.3	Before you start	75
3.2	Procedure	76
3.2.1	Using the Collocation option	76
3.2.2	Investigating collocates using the Sort option	78
3.2.3	Printing solutions	81
3.2.4	Investigating collocations without downloading: are men as handsome as women are beautiful?	82
3.3	Discussion and suggestions for further work	83

	3.3.1	The significance of collocations	83
	3.3.2	Some similar problems	85
4	A query too far		86
	4.1	The problem: variation in phrases	86
	4.1.1	A first example: 'the horse's mouth'	86
	4.1.2	A second example: 'a bridge too far'	86
	4.1.3	Highlighted features	86
	4.1.4	Before you start	87
	4.2	Procedure	87
	4.2.1	Looking for phrases using Phrase Query . . .	87
	4.2.2	Looking for phrases using Query Builder . .	91
	4.3	Discussion and suggestions for further work	95
	4.3.1	Looking for variant phrases	95
	4.3.2	Some similar problems	95
5	Do people ever say 'you can say that again'?		98
	5.1	The problem: comparing types of texts	98
	5.1.1	Spoken and written varieties	98
	5.1.2	Highlighted features	98
	5.1.3	Before you start	99
	5.2	Procedure	99
	5.2.1	Identifying text type by using the Source option	99
	5.2.2	Finding text-type frequencies using SGML Query	100
	5.2.3	Displaying SGML markup in solutions	103
	5.2.4	Searching in specific text-types: 'good heav- ens' in real and imagined speech	105
	5.2.5	Comparing frequencies in different text-types	107
	5.3	Discussion and suggestions for further work	108
	5.3.1	Investigating other explanations: combining attributes	108
	5.3.2	Some similar problems	110
6	Do men say mauve?		112
	6.1	The problem: investigating sociolinguistic variables . . .	112
	6.1.1	Comparing categories of speakers	112
	6.1.2	Components of BNC texts	112
	6.1.3	Highlighted features	114
	6.1.4	Before you start	115
	6.2	Procedure	115
	6.2.1	Searching in spoken utterances	115

6.2.2	Using Custom display format	116
6.2.3	Comparing frequencies for different types of speaker: male and female 'lovely'	117
6.2.4	Investigating other sociolinguistic variables: age and 'good heavens'	123
6.3	Discussion and suggestions for further work	125
6.3.1	Sociolinguistic variables in spoken and writ- ten texts	125
6.3.2	Some similar problems	126
7	'Madonna hits album' — did it hit back?	129
7.1	The problem: linguistic ambiguity	129
7.1.1	The English of headlines	129
7.1.2	Particular parts of speech, particular portions of texts	129
7.1.3	Highlighted features	130
7.1.4	Before you start	131
7.2	Procedure	131
7.2.1	Searching for particular parts-of-speech with POS Query: 'hits' as a verb	131
7.2.2	Searching within particular portions of texts with Query Builder	133
7.2.3	Searching within particular portions of par- ticular text-types with CQL Query	135
7.2.4	Displaying and sorting part-of-speech codes	137
7.2.5	Re-using the text of one query in another: 'hits' as a noun	138
7.2.6	Investigating colligations using POS collating	139
7.3	Discussion and suggestions for further work	140
7.3.1	Using part-of-speech codes	140
7.3.2	Some similar problems	141
8	Springing surprises on the armchair linguist	143
8.1	The problem: intuitions about grammar	143
8.1.1	Participant roles and syntactic variants	143
8.1.2	Inflections and derived forms	144
8.1.3	Variation in order and distance	145
8.1.4	Highlighted features	146
8.1.5	Before you start	146
8.2	Procedure	147

8.2.1	Designing patterns using Word Query: forms of 'spring'	147
8.2.2	Using Pattern Query	150
8.2.3	Varying order and distance between nodes in Query Builder	150
8.2.4	Checking precision with the Collocation and Sort options	153
8.2.5	Saving solutions with the Listing option . . .	154
8.3	Discussion and suggestions for further work	157
8.3.1	Using Two-way links	157
8.3.2	Some similar problems	158
9	Returning to more serious matters	161
9.1	The problem: investigating positions in texts	161
9.1.1	Meaning and position	161
9.1.2	Laughter and topic change	161
9.1.3	Highlighted features	162
9.1.4	Before you start	162
9.2	Procedure	163
9.2.1	Searching in sentence-initial position: 'anyway' and 'anyhow'	163
9.2.2	Customizing the solution display format . . .	165
9.2.3	Sorting and saving solutions in Custom format	168
9.2.4	Searching at utterance boundaries: laughs and laughing speech	170
9.3	Discussion and suggestions for further work	176
9.3.1	Searching near particular positions	176
9.3.2	Some similar problems	177
10	What does 'SARA' mean?	179
10.1	The problem: studying pragmatic features	179
10.1.1	How are terms defined?	179
10.1.2	Highlighted features	179
10.1.3	Before you start	180
10.2	Procedure	180
10.2.1	Looking for acronyms with POS Query . . .	180
10.2.2	Classifying solutions with bookmarks	181
10.2.3	Searching in specified texts	182
10.2.4	Serendipitous searching: varying the query type	184
10.2.5	Finding compound forms: combining Phrase and Word Queries	186

	10.2.6	Viewing bookmarked solutions	187
	10.2.7	Including punctuation in a query	189
10.3		Discussion and suggestions for further work	191
	10.3.1	Using Bookmarks	191
	10.3.2	Punctuation in different query types	191
	10.3.3	Some similar problems	192

III: Reference Guide 195

1		Quick reference guide to the SARA client	196
1.1		Logging on to the SARA system	196
1.2		The main SARA window	197
1.3		The File menu	199
	1.3.1	Defining a query	200
	1.3.2	Defining a Word Query	200
	1.3.3	Defining a Phrase Query	202
	1.3.4	Defining a POS Query	203
	1.3.5	Defining a Pattern Query	204
	1.3.6	Defining an SGML Query	206
	1.3.7	Defining a query with Query Builder	207
	1.3.8	Defining a CQL Query	210
	1.3.9	Execution of SARA queries	212
	1.3.10	Printing solutions to a query	213
1.4		The Edit menu	213
1.5		The Browser menu	215
1.6		The Query menu	216
	1.6.1	Editing a query	216
	1.6.2	Sorting solutions	217
	1.6.3	Thinning solutions	218
	1.6.4	Options for displaying solutions	218
	1.6.5	Additional components of the Query window	221
	1.6.6	Saving solutions to a file	222
	1.6.7	Displaying bibliographic information and brows- ing	223
	1.6.8	The Collocation command	224
1.7		The View menu	225
	1.7.1	Tool bar command	225
	1.7.2	Status bar command	226
	1.7.3	Font	227
	1.7.4	Colours	227
	1.7.5	User preferences	227

- 1.8 The Window menu 228
 - 1.9 The Help menu 229
 - 1.10 Installing and configuring the SARA client 229
- 2 Code tables 230
 - 2.1 POS codes in the CLAWS5 tagset 230
 - 2.2 Text classification codes 234
 - 2.3 Dialect codes 238
 - 2.4 Other codes used 239
- 3 SGML Listing format 240
- 4 Bibliography 242

