Unit 14 Language variation studies

14.1 Introduction

When people use language in different social and communicative contexts, their language often differs in terms of both grammatical and lexical choice. Biber et al (1999: 24) indicate that different registers or genres demonstrate consistent patterning. The authors find that many descriptions of general English, based on an averaging of patterns across registers, often obscure such register variation and are thus inaccurate and misleading. People who use the same language in different regions and countries may also talk differently. This unit presents four excerpts from published research in the area of language variation. The first extract, from Biber (1995a), provides an overview of Biber's framework of multi-feature/multi-dimensional (MF/MD) analysis (see also unit 10.4 and case study 5). In the second excerpt, Hyland (1999) undertakes a genre analysis on the basis of corpora of textbooks and research articles. The last two excerpts, Lehmann (2002) and Kachru (2003), are concerned with the regional variation of English.

14.2 Biber (1995a)

Biber and his colleagues have explored register and genre variation from three different perspectives: synchronic (e.g. Biber 1985, 1987, 1988), diachronic (e.g. Biber and Finegan 1989), and contrastive (e.g. Biber 1995b). Biber's MF/MD analysis framework has been well received as it establishes a link between form and function. This excerpt outlines the MF/MD approach and provides a background for case study 5 in Section C, which will compare conversation and speech in American English using Biber's approach and WordSmith Tools.

Biber, D. 1995a. 'On the role of computational, statistical, and interpretive techniques in multi-dimensional analysis of register variation'. *Text* 15/3: 314-370.

2. Overview of the multi-dimensional approach to register variation

The multi-dimensional approach to register variation was originally developed for comparative analyses of spoken and written registers in English (e.g., Biber, 1986, 1988). Methodologically, the approach uses computer-based text corpora, computational tools to identify linguistic features in texts, and multivariate statistical techniques to analyze the co-occurrence relations among linguistic features, thereby identifying underlying dimensions of variation in a language.

The primary research goal of the multi-dimensional approach is to provide comprehensive descriptions of the patterns of register variation, including (1) identification of the underlying linguistic parameters, or dimensions, of variation, and (2) specification of the linguistic similarities and differences among registers with respect to those dimensions. Two primary motivations for the multi-dimensional approach are the assumptions that: (1) generalizations concerning register variation in a language must be based on analysis of the full range of spoken and written registers; and (2) no single linguistic parameter is adequate in itself to capture the range of similarities and differences among spoken and written registers. The approach thus requires analysis of numerous spoken and written registers with respect to numerous linguistic features.

Some of the general characteristics of the multi-dimensional approach are:

1. It is corpus-based, depending on analysis of a large collection of naturally-occurring texts.

2. It uses automated computational techniques to analyze linguistic features in texts. This characteristic enables distributional analysis of many linguistic features across many texts and text varieties.

3. It uses interactive computational techniques to check the analysis of ambiguous linguistic features, ensuring accuracy in the final feature counts.

4. The research goal of the approach is the linguistic analysis of texts, registers, and text types, rather than analysis of individual linguistic constructions.

5. The approach is explicitly multi-dimensional. That is, it is assumed that multiple parameters of variation will be operative in any discourse domain.

6. The approach is quantitative. Analyses are based on frequency counts of linguistic features, describing the relative distributions of features across texts. Multivariate statistical techniques are used to identify co-occurrence patterns among linguistic features and to analyze the relations among texts.

7. The approach synthesizes quantitative and functional methodological techniques. That is, the quantitative statistical analyses are interpreted in functional terms, to determine the underlying communicative functions associated with each empirically determined set of co-occurring linguistic features. The approach is based on the assumption that statistical co-occurrence patterns reflect underlying shared communicative functions.

Dimensions represent distinct groupings of linguistic features that have been empirically determined to co-occur with significant frequencies in texts. It is important to note that the co-occurrence patterns underlying dimensions are identified quantitatively (by a statistical factor analysis) and not on any a priori basis. Dimensions are subsequently interpreted in terms of the communicative functions shared by the co-occurring features. Interpretive labels are posited for each dimension, such as ‘Involved versus Informational Production’ and ‘Narrative versus Non-narrative Concerns’.

In earlier synchronic multi-dimensional analyses of English (e.g., Biber 1986, 1988), approximately 500 texts from 23 registers were analyzed, including face-to-face conversations, interviews, public speeches, broadcasts, letters, press reportage, official documents, academic prose, and fiction. Subsequent analyses have used this approach to analyze texts from a number of more specialized registers, such as elementary school textbooks and student writing, job interviews, and the writings of individual authors. Linguistic features analyzed in these studies include both lexical and grammatical characteristics of texts (see section 2.3 below).

Individual texts, or groups of texts called registers, can be compared along each dimension. Two registers are similar along a dimension to the extent that they use the co-occurring features of the dimension in similar ways. Multi-dimensional analyses show that a pair of registers are often similar along one dimension (i.e., with respect to one set of co-occurring linguistic features) but quite different along another dimension (i.e., with respect to another set of features).

2.1. The use of automated and interactive computational techniques in multi-dimensional analyses

The use of automated and semi-automated (i.e., interactive) computational techniques is a practical rather than necessary aspect of multi-dimensional analyses. Such analyses by hand would be extremely time-consuming, and they are often considerably less reliable and accurate than analyses by computer.

Before the use of computers, empirical discourse analyses were typically based on a few thousand words of text; an analysis of 10,000 words was regarded as a major undertaking that required a long research period. Similarly it was possible to consider only a relatively restricted range of linguistic characteristics; analyses considering 10 different linguistic characteristics were regarded as major projects. In contrast, early multi-dimensional analyses employing computational techniques were based on a much more adequate and representative database: a text corpus over 100 times as large as in most previous analyses (nearly 1 million words of text), and inclusion of a very wide range of linguistic characteristics (67 different features in Biber [1988]).

Needless to say, some linguistic analyses must be checked interactively, because current automated techniques are not sufficiently accurate. For example, the distinction between some past tense verbs and past participial verbs functioning as post-nominal modifiers is notoriously hard for automated computer analyses. All automated grammatical taggers have difficulties dealing with distinctions such as this, and as a result, it is necessary to include interactive post-editing to insure accuracy (see section 3.2.2 below).

2.2. Methodological overview of the multi-dimensional approach

The multi-dimensional approach involves the following methodological steps:

1. Texts are collected, transcribed (in the case of spoken texts), and input into the computer. The situational characteristics of each spoken and written register are noted during data collection.

2. The published literature is reviewed, and if necessary supplemented by original grammatical research, to determine the range of linguistic features to be included in the analysis, together with functional associations of individual features (see, for example, Aijmer, 1984; Altenberg, 1984; Beaman, 1984; Chafe, 1982; Coates, 1983; Schiffrin, 1981, 1987; Tannen, 1982; Thompson, 1983; Tottie, 1986).

3. Computer programs are developed for automated grammatical analysis, to ‘tag’ all relevant linguistic features in texts.

4. The entire corpus of texts is tagged automatically by computer.

5. All texts are post-edited interactively to insure that the linguistic features are accurately identified.

6. Additional computer programs are developed and run to compute frequency counts of each linguistic feature in each text of the corpus.

7. The co-occurrence patterns among linguistic features (across all texts in the corpus) are analyzed, using a factor analysis of the frequency counts.

8. The co-occurrence patterns identified by the factor analysis are interpreted functionally as underlying dimensions of variation.

9. Dimension scores for each text are computed by summing the major linguistic features empirically grouped on each dimension; the mean dimension scores for each register are then compared to analyze the salient linguistic similarities and differences among spoken and written registers.

10. The functional interpretation of each dimension is revised based on the distribution of spoken and written registers along the dimension.

2.3. Choice of linguistic features included in multi-dimensional analyses

Although the co-occurrence patterns underlying dimensions are determined empirically, those patterns depend on the prior choice of linguistic features to be used in the analysis. Most multi-dimensional analyses to date have focused on lexical, grammatical, and syntactic features, with the goal of being as inclusive as possible. That is, any linguistic characteristic that can be interpreted as having functional associations is a candidate for inclusion in multi-dimensional analyses. Previous analyses have included:

– lexical features, such as type-token ratio and word length;

– semantic features relating to lexical classes, such as hedges, emphatics, speech act verbs, mental verbs;

– grammatical feature classes, such as nouns, prepositional phrases, attributive and predicative adjectives, past tense verbs, perfect aspect verbs, personal pronouns; and

– syntactic features, such as relative clauses, adverbial clauses, *that* complement clauses, passive postnominal participial clauses.

One characteristic of multi-dimensional analyses is that they can be extended by investigating the role of additional features in relation to previously determined dimensions. For example, Biber (1992b) analyzes the distribution and function of linguistic features marking reference and cohesion within texts, showing how these features relate to the previously identified multi-dimensional structure of English. While some cohesion features function as part of previously identified dimensions, other cohesion features co-occur in new patterns to define additional dimensions associated with the marking of reference in discourse.

Table 1. Summary of functions, linguistic features, and characteristic registers for the five major English dimensions identified in Biber (1988)

Functions	Linguistic features	Characteristic registers
Dimension 1 ‘Involved versus informational production’		
Involved (Inter)personal focus Interactive Personal stance On-line production	1st and 2nd person pronouns, questions, reductions, stance verbs, hedges, emphatics, adverbial subordination	Conversations, personal letters, public conversations
Informational Careful production Faceless	nouns, adjectives, prepositional phrases, long words	informational exposition, e.g., official documents, academic prose
Dimension 2 ‘Narrative versus non-narrative concerns’		
Narrative	past tense, perfect aspect, 3rd person pronouns, speech act (public) verbs	fiction
Non-narrative	present tense, attributive adjectives	exposition, broadcasts, professional letters, telephone conversations
Dimension 3 ‘Elaborated versus situation-dependent reference’		
Elaborated Situation-independent reference	WH relative clauses, pied-piping constructions, phrasal coordination	official documents, professional letters, written exposition
Situation-dependent reference On-line production	time and place adverbials	broadcasts, conversations, fiction, personal letters
Dimension 4 ‘Overt expression of persuasion’		
Overt argumentation and persuasion	modals (prediction, necessity, possibility), suasive verbs, conditional subordination	professional letters, editorials
Not overtly argumentative	–	broadcasts, press reviews
Dimension 5 ‘Abstract versus non-abstract style’		
Abstract style	agentless passives, by passives, passive dependent clauses	technical prose, other academic prose, official documents
Non-abstract	–	conversations, fiction, personal letters, public speeches, public conversations, broadcasts

Future multi-dimensional analyses could be extended to include linguistic features from additional domains, such as the frequency of various rhetorical devices or the frequency of different organizational patterns. Any text characteristic that is encoded in language and can be reliably identified and counted is a potential candidate for inclusion. Multi-dimensional analyses to date have focused primarily on a wide range of lexical and grammatical characteristics, but these analyses could be usefully extended to include consideration of language characteristics from other linguistic levels.

2.4. Summary of the 1988 multi-dimensional analysis of register variation in English

As noted in the introduction, it is important to distinguish between the multi-dimensional approach to register variation and multi-dimensional studies of particular discourse domains in particular languages. Watson focuses on the multi-dimensional analysis of English register variation presented in Biber (1988); this study provides the fullest account of multi-dimensional methodology and a synchronic analysis of the relations among adult spoken and written registers.

Five major dimensions are identified and interpreted in Biber (1988: especially chapters 6–7). Each comprises a set of co-occurring linguistic features; each defines a different configuration of similarities and differences among spoken and written registers; and each has distinct functional underpinnings. The five dimensions are interpretively labeled as follows:

1. Involved versus Informational Production
2. Narrative versus Non-narrative Concerns
3. Elaborated versus Situation-Dependent Reference
4. Overt Expression of Persuasion
5. Abstract versus Non-abstract Style

The primary communicative functions, major co-occurring features, and characteristic registers associated with each dimension are summarized in Table 1. Registers differ systematically along each of these dimensions, relating to functional considerations such as interactiveness, involvement, purpose, and production circumstances; and these functions are in turn realized by systematic co-occurrence patterns among linguistic features. The Appendix provides a more concrete illustration of how the 1988 multi-dimensional analysis can be used for comparative studies of spoken and written registers.

Two major conclusions come out of the 1988 multi-dimensional analysis of register variation in English: (1) no single dimension of variation is adequate in itself to account for the range of similarities and differences among registers — rather, multi-dimensional analyses are required; and (2) there is no absolute difference between spoken and written language rather, particular types of speech and writing are more or less similar with respect to different dimensions.

14.3 Hyland (1999)

Hyland (1999) compares the features of the specific genres of metadiscourse in introductory course books and research articles on the basis of a corpus consisting of extracts from 21 university textbooks for different disciplines and a similar corpus of research articles. This excerpt presents the methodology and findings of the paper.

Hyland, K. 1999. 'Talking to students: metadiscourse in introductory coursebooks'. *English for Specific Purposes* 18/1: 3-26.

Corpus and Procedure

The corpus consists of extracts from 21 introductory coursebooks in three academic disciplines: microbiology, marketing and applied linguistics, comprising almost 124 000 words (see Appendix A). The average length of the extracts was 5 900 words (range 3 305–10 678) consisting of complete chapters (16) or substantial sections of chapters beginning with the introductory matter and comprising entire contiguous sub-sections (5). The textbooks were selected from reading lists for introductory undergraduate courses and all extracts were among those recommended by teachers as containing 'core' reading matter. A parallel corpus of 21 research articles (121 000 words/average length 5 771 words) was compiled for comparison from the current issues of prestigious journals recommended by expert informants in the same three disciplines. The corpora were analysed independently by myself and two research assistants by coding all items of metadiscourse according to the schema outlined above. An interrater reliability of 0.83 (Kappa) was obtained, indicating a high degree of agreement.

Findings

Overall, the quantitative analysis revealed the importance of metadiscourse in these textbooks with an average of 405 examples per text; about one every 15 words. It should be noted here that the expression of devices according to a word count is not intended to represent the *proportion* of text formed by metadiscourse. Clearly, metadiscourse typically has clause-level (or higher) scope and I have standardised the raw figures to a common basis merely to compare the *occurrence*, rather than the length, of metadiscourse in corpora of unequal sizes. Table 2 shows that writers used far more textual than interpersonal forms in this corpus, and that connectives and code glosses were the most frequent devices in each discipline. The numerical preponderance of textual devices emphasises the common interpretation of metatext as guiding the reading process by indicating discourse organisation and clarifying propositional meanings.

TABLE 2 Metadiscourse in Academic Textbooks per 1 000 Words (% of total)

Category	**Biology**		**Applied Linguistics**		**Marketing**	
Logical connectives	32.3	(43.2)	17.8	(30.6)	34.4	(48.8)
Code glosses	9.4	(12.6)	9.6	(15.6)	9.7	(13.8)
Endophoric markers	6.4	(8.6)	4.5	(7.3)	2.5	(3.5)
Frame markers	2.5	(3.3)	4.6	(7.4)	4.2	(6.0)
Evidentials	3.2	(4.2)	5.3	(8.6)	1.0	(1.5)
Textual	**53.8**	**(71.9)**	**42.8**	**(69.4)**	**51.9**	**(73.7)**
Hedges	8.9	(12.0)	4.7	(7.7)	5.9	(8.4)
Emphatics	5.0	(6.7)	2.4	(3.9)	3.3	(4.7)
Attitude markers	4.1	(5.5)	3.5	(5.6)	5.5	(7.9)
Relational markers	2.2	(3.0)	6.1	(9.8)	2.5	(3.5)
Person markers	0.7	(0.9)	2.2	(3.6)	2.2	(3.6)
Interpersonal	**21.0**	**(28.1)**	**18.9**	**(30.6)**	**18.9**	**(30.6)**
Totals	**74.8**	**(100)**	**61.7**	**(100)**	**70.4**	**(100)**

TABLE 3 Ranked Metadiscourse Categories (Combined Disciplines)

	Textbooks		**Research articles**	
	Items per 1000 words	**% of total**	**Items per 1000 words**	**% of total**
Textual	**49.1**	**71.7**	**34.8**	**52.6**
Interpersonal	**19.4**	**28.3**	**31.4**	**47.4**
Subcategory				
Logical connectives	28.1	40.9	12.3	18.5
Code glosses	9.6	14.0	7.6	11.5
Hedges	6.4	9.4	16.7	25.3
Endophoric markers	4.4	6.5	3.2	4.9
Attitude markers	4.3	6.3	4.5	6.8
Frame markers	3.8	5.5	5.6	8.5
Relational markers	3.7	5.4	2.5	3.8
Emphatics	3.5	5.1	4.2	6.3
Evidentials	3.3	4.8	6.1	9.3
Person markers	1.4	2.1	3.5	5.2
Grand Totals	**68.5**	**100%**	**66.2**	**100%**

TABLE 4 Metadiscourse in Textbooks and RAs per 1 000 Words

	Biology		**Applied Linguistics**		**Marketing**	
	Textbook	**RA**	**Textbook**	**RA**	**Textbook**	**RA**
Textual	53.8	40.1	42.8	30.1	51.9	36.6
	71.9%	66.8%	69.4%	49.2%	73.7%	49.7%
Interpersonal	21.0	19.9	18.9	31.0	18.5	37.0
	28.1%	33.2%	30.6%	50.8%	26.3%	50.3%
Totals	74.8	59.9	61.7	60.1	70.4	73.6

TABLE 5 Proportions of Metadiscourse in RAs and Textbooks

	Biology		**Applied Linguistics**		**Marketing**	
Category	**TB**	**RA**	**TB**	**RA**	**TB**	**RA**
Logical connectives:	43.2	18.8	30.6	18.1	48.8	18.7
Frame markers	3.3	8.6	7.4	7.6	6.0	9.0
Endophoric markers	8.6	7.7	7.3	4.1	3.5	4.4
Evidentials	4.2	16.2	8.6	7.3	1.5	8.0
Code glosses	12.6	15.4	15.6	12.1	13.8	9.6
Textual	**71.9**	**66.8**	**69.4**	**49.2**	**73.7**	**49.7**
Hedges	12.0	20.0	7.7	25.6	8.4	27.0
Emphatics	6.7	5.8	3.9	7.4	4.7	5.7
Attitude markers	5.5	2.2	5.6	8.8	7.9	7.0
Relational markers	3.0	1.2	9.8	4.1	3.5	4.5
Person markers	0.9	4.0	3.6	4.8	1.8	6.0
Interpersonal	**28.1**	**33.2**	**30.6**	**50.8**	**26.3**	**50.3**
Total %	100	100	100	100	100	100

The tables show some obvious disciplinary variations in metadiscourse use. The applied linguistics texts comprise considerably more evidentials and relational markers, the biology authors favoured hedges, and marketing textbooks had fewer evidentials and endophorics. Perhaps more interesting however are the cross-discipline similarities, with all three fields containing comparable total use and a near identical proportion of textual and interpersonal forms. In particular, all disciplines showed a high use of logical connectives and code glosses which together comprised about half of all cases, demonstrating that the principal concern of textbook authors is to present information clearly and explicitly.

A comparison with the research articles revealed strikingly similar total frequencies of metadiscourse in the two corpora, but a considerable difference in the proportion of the two main categories (Table 3). The increase in interpersonal metadiscourse from about a third of all cases in the textbooks to nearly half in the RAs shows the critical importance of these forms in persuasive prose.

As can be seen, devices used to assist comprehension of propositional information, such as connectives, code glosses and endophoric markers, were less frequent in the articles while those typically used to assist persuasion, such as hedges, emphatics, evidentials and person markers, were more frequent. Hedges were almost three times more common in the RAs and represented the most frequent metadiscourse feature, demonstrating the importance of distinguishing established from new claims in research writing and the need for authors to evaluate their assertions in ways that their peers are likely to find persuasive.

When separating the texts by both discipline and genre we find that the tables above mask a number of variations in metadiscourse use. Table 4 shows that the overall density levels differed markedly in biology, with almost 25% more metadiscourse in the textbooks than the RAs, due mainly to a heavier use of textual forms. Biology was also the only discipline where there was little change in the proportions of interpersonal and textual features between the two genres, while the interpersonal frequencies increased dramatically in the applied linguistics and marketing RAs.

Table 5 shows that the use of logical connectives was highest in textbooks in all disciplines and that the RAs contained a higher proportion of hedges, person and frame markers. Biologists showed the greatest variation, both across genres and disciplines, with substantial genre differences in most categories. While the marketing and applied linguistics texts were more uniform between genres, both contained large differences in hedges and connectives. Substantial genre variations were also apparent in the use of evidentials and person markers in marketing and endophoric and relation markers in applied linguistics. In general, metadiscourse variations were more pronounced between genres than disciplines, particularly for high frequency items, and the textbooks tended to exhibit greater disciplinary diversity than the RAs.

Discussion

Textbooks, as a specific form of language use and social interaction, both represent particular processes of production and interpretation, and link to the social practices of the institutions within which they are created. We might expect, then, that metadiscourse variations will reflect the different roles that textbooks and research papers play in the social structures of disciplinary activity and anticipate that their use will contain clues about how these texts were produced and the purposes they serve. Metadiscourse is grounded in the rhetorical purposes of writers and sensitive to their perceptions of audience, both of which differ markedly between the two genres. One audience consists of an established community of disciplinary peers familiar with the conceptual frameworks and specialised literacies of their discipline. The other is relatively undifferentiated in terms of its experience of academic discourse, often possessing little more than a general purpose EAP competence in the early undergraduate years (e.g. Leki & Carson 1994). As a result of such contextual differences, what can be said, and what needs to be said, differs considerably. It is therefore interesting to speculate on the patterns observed and I will consider textual and interpersonal variations in turn.

14.4 Lehmann (2002)

Lehmann (2002) presents a large-scale study of zero-subject relative constructions (ZSRs) on the basis of the demographically sampled spoken BNC and the five-million-word Longman Spoken American Corpus (LSAC). This study shows that there is a sharp difference between American English which has 2.5% subject relatives with a zero relativizer and British English which has 13%.

Lehmann, H. 2002. 'Zero subject relative constructions in American and British English'. *New Frontiers in Corpus Research*, pp. 163-177. Amsterdam: Rodopi.

6. Results

The analysis left me with 94 instances of ZSRs in 5 million words in LSAC and 205 instances in 4.2 million words in the spoken demographic sample of the BNC. This certainly is a strong indication that ZSRs are about two and a half times more frequent in British English than in American English. However, as a consequence of the principle of accountability there are problems with accounting frequency per million running words. After all it could be the case that subject relative constructions allowing for a possible realization by zero are less frequent in American English overall. For this reason it is important to account for the frequency of ZSRs taking into account the overall frequency of possible occurrences including surface relativizers. The results of such an analysis are shown in Table 1.

Table 1 shows that taking into account occurrences and non-occurrences of ZSRs does not reduce the difference found above. In fact, taking into account possible occurrences results in an even greater difference between American and British English, with ZSRs being over five times more frequent in British English than in American English. Thus the results in Table 1 firmly establish a pronounced difference in the use of ZSRs between these two major varieties of English.

Table 1: Realization forms of subject relatives in American and British English

American English				**British English**			
surface		**zero**		**surface**		**zero**	
N	%	n	%	n	%	n	%
3647	97.5%	94	2.5%	1376	87%	205	13%

Table 2: Types of ZSRs in American and British English

	American English		**British English**	
Types	**n**	**%**	**n**	**%**
existential *there*	27	29%	126	61%
Cleft	24	26%	25	12%
Be	14	15%	8	4%
Have	9	10%	15	7%
Others	20	21%	31	15%

Table 1 also shows a striking difference in the frequency of subject relative constructions observed by the retrieval patterns. Subject relative constructions conforming to the retrieval patterns are twice as frequent in American English as in British English.

Another interesting aspect for analysis is the different matrix clauses in which ZSR complexes occur. Here I will follow Shnukal (1981), who defines four major types exemplified by (22)–(25).

(22) there's this woman Ø went out to like some Caribbean or something to have a vacation and then she met this guy (LSAC: 118901:?)

(23) It was Joanne Ø said you'd go down there, so you said alright. (BNC:KDG: 1795:PS000)

(24) this dog I got friendly with, they were people Ø got in there for the summer, got and just abandon it so there's all these dogs running around the coast. (LSAC:165301#dr2791)

(25) well I mean we had one girl Ø didn't know what she was going about, … (BNC:KDW:6217:PS1C1)

The first type represented by (22) is characterized by existential *there* in the matrix clause. The second type features an *it*-cleft construction, as in (23). The third type has the verb *be* as a main verb, as in (24). The fourth type of matrix clause is characterized by the verb *have* as a main verb, as in (25). These are the most frequent matrix clause types discussed by Shnukal (1981). There are other less frequent verbs used in the matrix clauses in my material as in (26)–(31).

(26) … and erm thought knock on the doors and se, ask people Ø had seen it and (BNC:KD5:9674:PS0JX)

(27) knife at the back of the saw, they, it is a bit dangerous, erm where's the guard Ø goes at the top. (BNC:KDM:07067:PS0RD)

(28) And I can handle any bastard Ø gets in here. (BNC:KDY:0658:PSI42)

(29) and it was talking in there <u>about one woman Ø asked the waiter to go to bed with him when she was ordering something</u>. (LSAC:151103:2173)

(30) It's like people who eat marmalade Ø has no peel in it. (BNC:KPU:2690:PS584)

(31) either you kill each other until finally you get to one Ø gives in so one becomes the master and one becomes the slave. (LSAC:165001:2757)

(26) – (31) show that the construction is not limited to the matrix clause types exemplified in (22) – (25). Table 2 shows the distribution of the four main types and other matrix clause choices.

Table 2 shows that the ranking of matrix clause types in American and British English is fairly similar. The only exceptions are constructions with the verb *be* and the verb *have* which are reversed in the order of their ranking. The most striking difference is found with existential *there* constructions, which cover 61% of all ZSR constructions in British English and only 29% in American English.

In the following I will try to analyze the distribution of the ZSR construction with the help of the annotation provided by the BNC and the LSAC. Both corpora are annotated with social variables for individual speakers. However, not all speakers are annotated, as it was impossible to obtain the relevant data from all interlocutors who happened to be recorded. The use of speaker annotation divides the corpora into parts of unequal size. To cope with this problem I prepared databases containing annotation and word-counts for the individual speakers. This information was then used for normalization and thus made direct comparison possible.

Table 3: Distribution of ZSRs over ethnic groups in spoken American English

Ethnicity	**No of Words**	**ZSR**	**ZSRs per 1**
		(n)	**million words**
White	2,493,493	55	22
Not Indicated	800,820	34	42
Black/African-American	145,989	4	27
Other	84,474	1	12
Multiple/Mixed ethnicity	67,520	–	–
Chicano/Mexican-American	59,884	–	–
American Indian/Native American	56,280	–	–
Latino/Other Hispanic	54,893	–	–
Filipino/Filipino-American	22,967	–	–
Polynesian/Pacific Islander	13,424	–	–
Chinese/Chinese-American	11,397	–	–
Korean/Korean-American	10,874	–	–
Japanese/Japanese-American	8,319	–	–
Puerto Rican	1,446	–	–
Arab/Arab-American	404	–	–
Viet/Thai/Other Asian	347	–	–

Table 4: Distribution of ZSRs in the LSAC according to age of speaker

age of speaker	**n**	**Number of words**	**Freq. per 1 m words**
0–14	1	46246	21.6
15–24	11	852081	12.9
25–34	18	803259	22.4
35–44	7	694733	10.0
45–59	19	889398	21.4
60+	10	295592	33.8

LSAC contains information about the ethnicity of speakers. Given the presence of ZSRs in AAVE [African American Vernacular English] documented in Tottie and Harvie (1999) and Harvie (1998), it is interesting to see if African Americans use ZSRs with a higher frequency than other Americans. Table 3 highlights the problem of correlating social variables with a low frequency phenomenon like ZSRs. The slightly higher frequency of ZSRs produced by African Americans is undermined by the fact that it is based on only four instances. The conclusion that can be drawn from Table 3, nevertheless, is that the majority of the occurrences are produced by European Americans and African Americans. The presence of ZSRs in spoken American English can thus not be attributed to the language use of one single ethnic group like AAVE. The absence of ZSRs from all other ethnic groups is certainly noteworthy. The high frequency of ZSRs for which speaker ethnicity is not indicated certainly raises the question of the observer's paradox. Speakers for which there is no annotation available are likely to be passers by. They may therefore be less conscious of being recorded.

Another interesting variable is the age of speakers presented in Table 4.

Here again no clear picture emerges. The highest frequency for the group 60+ might suggest that ZSRs are used more frequently by older speakers. However, the data for the other age groups doesn’t support such a trend. The gender of speakers using ZSRs is remarkably even: female speakers with 19 instances per million words, and male speakers with 18 instances per million words. Nor is there any support for the hypothesis that speakers using ZSRs belong to a lower social class. In terms of occupation we find professors, graduate students, lawyers, bankers as well as a cashier and a seamstress who produce instances of ZSRs.

The analysis of social variables for the British data was more successful. This is to be expected given the higher number of instances available for a breakdown according to the individual social variables, firstly on speaker age. Table 5 shows the frequency of ZSRs according to age of speakers.

Table 5 and its graphical representation in Figure 1 show a clear increase of the use of ZSRs with increasing age of speaker. However, such a result might be attributed to a difference in frequency of all the subject relatives under observation and not only to zero realizations. Table 6 shows the proportion of SSR [surface subject relative] and ZSR constructions according to age of speaker.

Table 5: Distribution of ZSRs in the BNC according to age of speaker

Age of speaker	**n**	**frequency per 1 m words**
0–14	8	19.2
15–24	9	21.5
25–34	25	36.3
35–44	36	51.3
45–59	51	72.3
60+	49	74.2
not available	27	42.9

Table 6: Proportion SSRs and ZSRs according to age of speaker in the BNC

	surface		**zero**	
age of speaker	**n**	**%**	**n**	**%**
0–14	95	92.2	8	7.8
15–24	114	92.7	9	7.3
25–34	223	89.9	25	10.1
35–44	248	87.3	36	12.7
45–59	246	82.8	51	17.2
60+	273	84.7	49	15.3

Table 6 shows that the trend observed in Table 5 and Figure 1 can’t be attributed to subject relatives in general. Zero as a relativizer choice is still twice as frequent in the age group 60+ than in the age group 0–14. Figure 2 shows a graphical representation of the frequency information in Tables 4 and 5. It helps us to visualize both the distribution of subject relatives in general and the distribution of surface and zero variants.

While the younger speakers certainly use fewer subject relatives, this cannot explain away the trend found in the distribution of ZSRs. This is documented by the proportion of realizations by zero, which increases with the age of speaker. Shnukal (1981: 322) comes to a similar conclusion for a dialect of Australian English.

The most important findings presented above are the difference between American English and British English and the fact that younger speakers use fewer ZSRs than older speakers. Given that, synchronically, language change is only observable as variation, a variation phenomenon like the realization form of subject relatives lends itself to speculation about an ongoing language change. On the assumption that language acquisition is completed at a certain age between 16 and 25, we may conclude that we are indeed observing an ongoing language change in British English, with an observable decrease of the use of ZSRs from the older to the younger generation. This is particularly important because variation and change do not mutually imply each other. While it is sound to extrapolate from ongoing language change to the presence of variation, the reverse does not hold. From observation of variation we may not extrapolate to the presence of language change. Variation phenomena can represent differences in registers, which may remain stable over time. Thus speaker age presents the only means of observing language change in synchronic corpora.

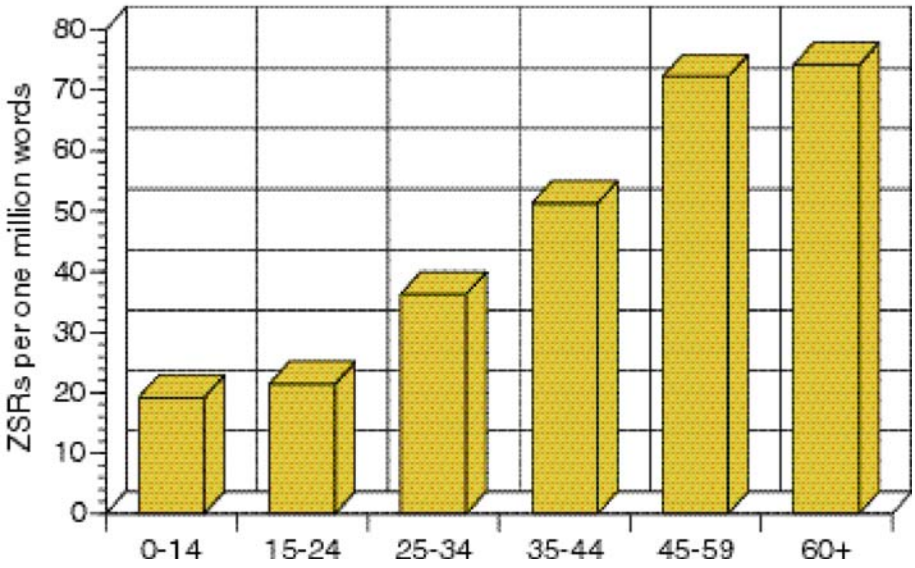

Figure 1: Distribution of ZSRs in the BNC according to age of speakers

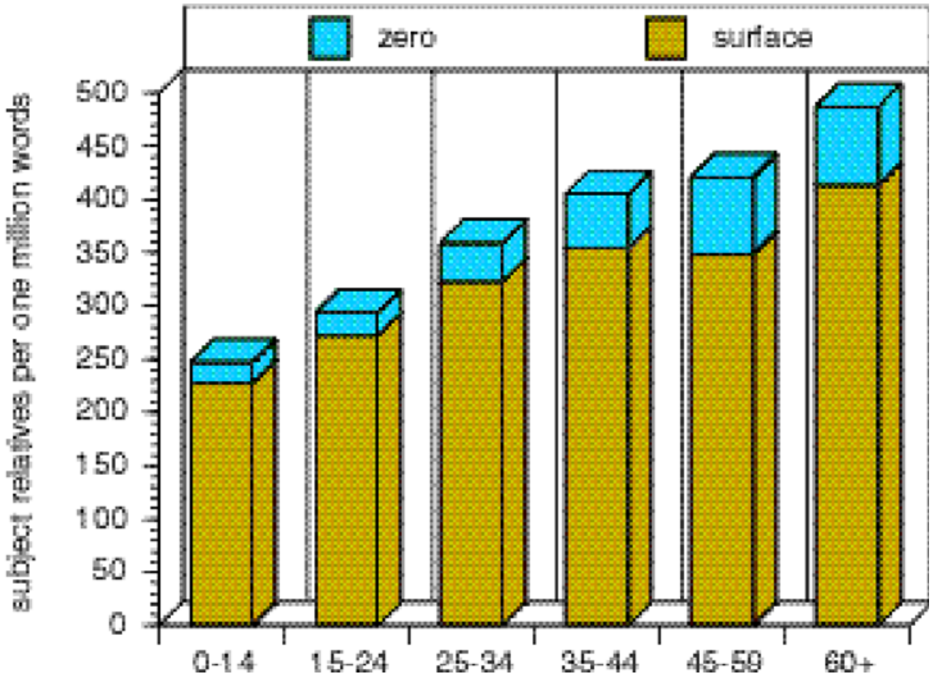

Figure 2: ZSRs and SSRs per million words in the BNC

However, using speaker age for documenting language change is certainly not uncontroversial. Only under the assumption that language use remains stable after the phase of language acquisition can speaker age be used for documenting ongoing language change. Even if language acquisition is completed at the age of about 18, this does not necessarily mean that the frequency of use of the acquired repertoire remains stable over the course of an adult's life.

The finding that ZSRs are over five times more frequent in British English than in American English could mean that the same language change – loss of ZSRs – has progressed further in American English. This may have been caused by the large number of immigrants speaking Western European languages like German, French, Italian and Spanish, which only have overt relativizers.

14.5 Kachru (2003)

Kachru (2003) uses a small corpus to explore the uses of definite reference across four regional varieties of English: Indian, Nigerian, Singaporean and American. The study indicates that the use of definite descriptions is likely to differ in 'Englishes' used in different parts of the world.

Kachru, Y. 2003. 'On definite reference in world Englishes'. *World Englishes* 22/4: 497-510.

The corpus collected for the study consisted of a number of letters to the editor from several newspapers in India, Nigeria, Singapore and the USA. The letters were published between March 5 and April 6, 2000. The total corpus was just over 15,000 words which yielded 945 noun phrases with the definite article *the*. The rationale for choice of data source was that newspapers correspond to a more casual style of writing and within newspapers the letters to the editor represented the least edited and most typical of individual style.

The classification of definite NPs in Quirk *et al.*, cited in Table 1, though useful in the classroom devoted to pedagogical grammar of English at a university setting, was too sketchy to serve the purposes of an analysis of definite reference in a larger corpus with the focus I have already mentioned. Although Quirk *et al.* (1985) also discuss several other categories of nouns with definite and indefinite articles, the classification suggested in Poesio and Vieira (1998) seemed more suitable for my purposes since it was developed for large-scale corpus analysis. Poesio and Vieira ran two experiments to determine how 'good' (i.e., how much they agree among themselves about analyzing definite descriptions) naive subjects are at doing the form of linguistic analysis presupposed by current schemes for classifying definite descriptions. Their subjects were asked to classify definite descriptions found in a corpus of natural language texts according to classification schemes developed starting from the taxonomies proposed in Hawkins (1978) and Prince (1981, 1992). The experiments were also designed to assess the feasibility of a system to process definite descriptions on unrestricted text and to collect data that could be used for their implementation.

I coded the noun phrases in my corpus according to the classification of noun phrases with the definite article *the* given in 1–6 in A below. This still left some definite NPs encountered in the corpus unaccounted for. I therefore had to add the classes 7–9 to the list of categories. There are still problems with the classification, which I will mention toward the end.

A. Classification of definite NPs

1. Anaphoric (definite NPs that cospecify with a discourse entity already introduced in the discourse).

John bought a car. ***The car/vehicle*** *turned out to be a lemon*.

2. Immediate situation (definite **NP** used to refer to an object in the situation of utterance; it may be visible or inferred).

At the dining table: *Please pass* ***the salt!***

Sign at the zoo: *Don't feed* ***the bears!***

3. Larger situation (in which the speaker appeals to the hearer's knowledge of entities that exist in the non-immediate or larger situation of utterance knowledge that speakers and hearers share by being members of the same community).

On a specific campus, talking about lunch: *Shall we meet in* ***the ballroom?***

4. Associative anaphoric (speaker and hearer may have (shared) knowledge of the relations between certain objects (the triggers) and their components and attributes (the associates); associative anaphoric use exploits such knowledge).

There was an accident at the intersection. ***The car*** *was smashed, but* ***the passengers*** *and* ***the driver*** *escaped without serious injury.*

5. Unfamiliar (definite NPs that are not anaphoric, do not rely on information about the situation of utterance, and are not associates of some triggers in the previous discourse).

(a) NP complements

the fact/suggestion *that …,* ***the place*** *where …*

(b) Nominal modifiers

the color *maroon,* ***the number*** *three*

(c) Referential relative

The book *that you were reading …*

(d) Associative clause (definite NPs that specify both the trigger and the associate)

The OP ED page *of the NY Times …*

(e) Unexplanatory modifiers

The last person *to leave the party was an old woman.*

6. Institutional (‘sporadic reference’ in Quirk *et al.*, 1985)

The USA, The UN

7. Fixed collocations (‘the logical use of the’ in Quirk *et al.*, 1985)

the first flight *to Denver, … catch* ***the last bus*** *…*

8. Generic

the musk ox, the tiger *…*

9. Idioms

a shot in ***the arm***

All the noun phrases with the definite article *the* in the corpus were coded in terms of the categories listed in 1–9 in A above. I have given one example of a text with coding in the Appendix to show how the analysis was done. I am not giving all the data on each piece of text and even each variety, mainly in view of space considerations.

The distribution of forms according to the classification in A is in Table 4. Although there are more occurrences of definite NPs of classes 1, 3, 4, and 5d, all anaphoric in some sense, the numbers still do not represent the anaphoric use as the most frequent use of definite NPs. This is clear from the percentages involved (Table 5).

The direct anaphoric referential NPs constitute under four tenths of the total use of definite NPs. Even if we combine the direct and indirect anaphoric referential use (i.e., classes 1 and 4), the percentage is still just about half of the total. Classes 3 and 5c are referential, but not anaphoric, and class 5d performs a deictic rather than a referential function. The other classes have a purely grammatical function. The occurrences in other categories are so small that it is hard to even speculate about their significance.

DISCUSSION AND INTERPRETATION

What is interesting about Tables 4 and 5 is that they do not point to a great deal of difference across varieties. This strengthens the claims that institutionalized varieties of English are not the same as learner varieties referred to by the term interlanguage. They are not ‘interference varieties’, as Quirk *et al.* (1985) characterize them.

Table 4. Distribution of definite NPs across varieties

	Classes											
	1	**3**	**4**	**5a**	**5b**	**5c**	**5d**	**5e**	**6**	**7**	**8**	**9**	**Total NPs**
USA E	75	20	27	7	6	12	34	1	11	2	5	2	202
IndE	101	55	26	8	1	7	66		6	11	5	5	291
S’Pore E	84	26	28	14	5	8	37	1	5	7	2	2	219
Niger E	84	13	39	7	1	10	43		13	15		2	227

Table 5. Percentages of NPs in classes across varieties

	USA E	Ind E	S'pore E	Niger E
1	37	35	38	37
3	10	19	12	6
4	13	9	13	17
5d	17	22	17	19
1 + 4	50	43	50	54

The results, however, raise a question about my third observation that in view of the descriptions of Asian and African varieties of English, it is reasonable to assume that world Englishes will exhibit differences in their use of definite NPs. Do the results presented here render this assumption invalid? I would submit that it is difficult to come to any conclusion on the basis of the small corpus I have analyzed so far. We need to have a much larger corpus with many more different types of texts to determine if inter-variety differences exist.

There is, however, another possibility. The data documented in earlier descriptions have not paid much attention to the cline of bilingualism in English (B. Kachru, 1965: 393–6). If we take the acrolectal varieties of English, there may be very little difference across them. As the users of the institutionalized varieties of English gain proficiency in the language and become users of the acrolectal variety, their cognitive abilities obviously make it possible for them to perceive and conceptualize grammatical structures which do not operate in their substratum languages. This is true of an entire range of grammatical phenomena. There is no reason why this should not be true also of the deictic and anaphoric relations and their exponents, which would explain their use of definite NPs similar to the Inner Circle users. The non-Inner Circle users' internalizing of the system of definite NPs in English is not precluded by their other language experience. This is an empirical question and needs to be investigated.

14.6 Unit summary and looking ahead

This unit was concerned with language variation. It first introduced Biber's MF/MD approach to register and genre analysis. The other three excerpts in this unit explored variation in specific genres and language varieties. In case study 5 in Section C of this book, we will compare Biber's analytic framework with an approach that uses WordSmith, which is less technically demanding and can approximate a Biber-style analysis. We will also, in case study 2 in Section C, consider the differences between British and American English, the two major varieties of English. While the excerpts presented in this unit are synchronic studies of a single language, we will explore language variation from contrastive and diachronic perspectives in the next unit.