シンポジウム

Current Trends in Corpus Linguistics: Voices from Britain

Laurence Anthony, Yasunori Nishina, Kaoru Takahashi, and Michael Handford

Abstract

This paper describes past and present research at three major corpus linguistics institutions in Britain, with the aim of providing a unique insight into some of the guiding forces that have helped shape the field of corpus linguistics over the past fifty years. Works carried out at the University of Birmingham, Lancaster University, and the University of Nottingham will be reviewed and their implications discussed. In addition, some tentative predictions about the future of corpus linguistics at these three institutions will be presented.

1. Introduction

Since the birth of modern corpus linguistics in the 1960s, innovative research has continued to emerge from leading British institutions, with the University of Birmingham, Lancaster University, and the University of Nottingham making some of the most significant contributions to the field (McEnery and Hardie, 2012: 73). At the University of Birmingham, for example, the COBUILD project, led by John Sinclair, resulted in the creation of the first corpus-informed learner dictionary, *Collins COBUILD English Language Dictionary* (Sinclair, 1987). The influential work of Tim Johns on data driven learning also emerged from the University of Birmingham, as has the pattern grammar work of Susan Hunston, and a host of other research projects. Similarly, Lancaster University has been home to many important developments in corpus linguistics. For example, much of the tagging work on the British National Corpus was carried out under the leadership of Geoffrey Leech at the Lancaster University Centre for Computer Corpus Research on Language (UCREL). Lancaster University is also home to leading researchers on language change, the language of minorities, and critical discourse studies (e.g., Baker, 2010). In the area of spoken language, many key findings have emerged from the CANCODE project at the University of Nottingham under the guidance of Michael McCarthy and Ronald Carter. The University of Nottingham is also

『英語コーパス研究』第19号（2012），pp.67-92

unique in its attempt to apply corpus linguistics techniques in the field of health and medicine.

In this paper, we will look at the history and research carried out at these three institutions in order to identify some of the major trends in British corpus linguistics that have shaped research in the field as a whole. First, we will attempt to address the intriguing question of why so many contributions to the field of corpus linguistics have emerged from British institutions. Then, in Sections 3 to 5, we will review the work at Birmingham, Lancaster, and Nottingham. Finally, we will conclude with a discussion on the similarities and differences between the three approaches to corpus research before finishing with a brief look at how these institutions have influenced corpus research in Japan.

2. Understanding the success of corpus linguistics in Britain

There is no doubt that British institutions have made very significant contributions to the field of corpus linguistics. Perhaps the most obvious reason for this is that British institutions in general have maintained a long and successful tradition of research in both the humanities and the sciences. University faculty in British institutions are strongly encouraged to pursue innovative research projects that receive highly competitive grants. In many cases, external grants may be the only form of funding that an individual receives. Also, promotion within a department will be largely determined by the output that a faculty member produces in terms international refereed research papers and successful grant awards. In addition, a faculty member will have minimal teaching loads compared to faculty in equivalent positions in other parts of the world, in particular, those working in Japan. Faculty are also likely to have fewer meetings, less administration responsibilities, and many more opportunities to collaborate with like-minded researchers around the world using English as a lingua franca. In the area of corpus linguistics, in particular, many of the early researchers were able to meet and discuss research through the International Computer Archive of Modern English (ICAME) organization that was established in the 1970s. ICAME is still a very influential organization in corpus linguistics, holding regular international conferences and hosting many key corpora.

There are other reasons for the success of corpus linguistics in Britain. Geoffrey Leech of Lancaster University considers that one of the critical factors for early successes was an acceptance of alternative theories to generative grammar, which dominated much of the work in the United States at the time (Nunez, 2006: 154). In the same paper, he also explains that from an early stage British publishers

were interested in the practical applications of corpus research, especially in the area of dictionaries. Clearly, the funding that corpus projects received from these commercial organizations had a positive influence not only on the number and size of corpus projects that could be initiated, but also on the number of people who were willing to take part in this new area of study.

McEnery and Hardie (2012) offer another perspective on this question. Although the first large-scale corpus that was available to researchers emerged from work in the United States, i.e., the Brown Corpus[1], McEnery and Hardie argue that by the 1990s British researchers had access to many more large, high-quality corpora, including the Bank of English[2], the British National Corpus[3], and the London-Lund Corpus[4]. In contrast, researchers interested in US English, or indeed any other language, had many fewer resources.

One final reason for the success of British corpus linguistics research is that British researchers have always had at their disposal easy-to-use corpus analysis tools, thanks to the efforts of technical astute researchers. For example, in the early days of corpus work Al Reed at the University of Birmingham developed the CLOC software that was used in the pioneering research that later developed into the COBUILD project (for details see Sinclair et al., 2004). Tim Johns, also at the University of Birmingham, developed the Micro-Concord concordancer (Johns, 1986), which could run on a modest home computer, such as the Spectrum. More recently, Mike Scott at the University of Liverpool, who worked with Tim Johns on later versions Micro-Concord[5], has continually developed the WordSmith Tools program[6], which is one of the most commonly used tools by both corpus linguists and also researchers in English for Specific Purposes (Hewings, 2011). More recent additions include AntConc (Anthony, 2011), which is the first multiplatform, standalone, concordance tool, also developed by a graduate of the University of Birmingham, and the BNCweb[7], developed at Lancaster University.

3. Corpus Linguistics at the University of Birmingham

3.1 The birth of corpus linguistics at the University of Birmingham

The founding of the Birmingham school of corpus linguistics can be traced back to an early corpus project entitled *English Lexical Studies* that was initiated at Edinburgh University in 1963 under the guidance of John Sinclair, and completed in 1970 after Sinclair had moved to the University of Birmingham. A full report of the project was later published as *English Collocation Studies: The OSTI Report* (Sinclair et al., 2004).

Sinclair's group was perhaps the first to conduct a lexical investigation based on a

corpus. During the six years of the *English Lexical Studies* project, they established many of terms, concepts, investigative procedures, and common practices that are still in use today. For example, the team introduced the terms 'span,' 'node,' and 'collocate.' They also stressed the importance of high frequency words and highlighted the important role of collocation, based on the fundamental ideas put forth by Palmer and Hornby (1933) and Firth (1957). They also developed the commonly used rule-of-thumb of using a four or five word span to establish the limits for collocation strength. Interestingly, although Birmingham is more associated with textual data, the *English Lexical Studies* project was concerned with 'messy' spoken data, and this led to one of the principal findings of the study, i.e., that grammar and lexis cannot and should not be treated separately (Sinclair et al., 2004: xviii).

After the publication of the OSTI report, corpus research at the University of Birmingham began to flourish. In the 1970s, Sinclair's group began to develop major corpora, including the 17 million-word *Birmingham Collection of English Text* in the 1980s and the 300 million-word *Bank of English* monitor corpus in the 1990s, which currently stands at over 450 million words. Sinclair's group also launched the COBUILD project with the publisher Collins (now HarperCollins), which led to the first corpus-informed learner corpus, *Collins COBUILD English Language Dictionary* (Sinclair, 1987), as mentioned in the introduction. Corpus-informed lexicography has continued to be a major strand of research at Birmingham, as seen in the continued publication of 2nd, 3rd, 4th, and 5th editions of the COBUILD dictionary and works such as Sinclair's (1991) seminar publication *Corpus, Concordance, Collocation.* Indeed, Sinclair's theoretical views and innovative approach to grammar and lexis are still a major influence on research at Birmingham, even though he moved to Italy in 2000 and sadly passed away in 2007.

3.2 Traditional concepts in corpus linguistics at the University of Birmingham

One of the core concepts in Sinclair's early work is a view of word meaning that is similar to that expressed by Firth (1935: 37): 'the complete meaning of a word is always contextual, and no study of meaning apart from a complete context can be taken seriously.' Sinclair's later works elaborate on the idea of 'unit of meaning' and further explain the ways in which meaning cannot belong to individual words but rather how word meaning is always defined by its co-texts (Sinclair 1991, 2004). Today, these ideas have been developed further by Susan Hunston and Wolfgang Teubert not only in the area of research, but also teaching. Teubert (2003: 9), for instance, points out that 'no word has a meaning except when it is encountered in context,' and Hunston (2011a: 14) assumes that 'the meaning of

any word cannot be identified reliably if the word is encountered in isolation.'

The idea that 'a word is always contextual' is comprehensively illustrated in Sinclair's corpus-based studies on lexical items (Sinclair, 1991; Stubbs, 2001). To identify 'meaning units,' many of Sinclair's studies prioritize the observing of sequences of words of varying degrees of fixedness with flexible boundaries as well as identifying the semantic similarity in the co-texts of a target word and phrase. 'Meaning units' can be realized through collocation between words (e.g., *food* + *assistance*) and colligation between words and grammatical categories (e.g., ADJ + *about* / ADJ + PREP). Other researchers have confirmed that texts are constructed from such sequences. For example, Hunston & Francis (1999) presented numerous ways in which lexis and grammar are closely linked as 'patterns,' and Erman & Warren (2000) showed that lexical-grammar 'sequences' account for 55.38% of whole texts.

Corpus linguists at Birmingham have further revealed that 'meaning units' are imbued by semantic and discoursal functions and that they express the writer's stance in texts (e.g., Charles, 2004, 2006a, 2006b; Hunston & Sinclair, 2000). Such discoursal features of a word (or sequence of words) are termed as semantic preferences between words and lexical sets (e.g., *in* + TIME) and semantic prosody that expresses 'the consistent aura of meaning' of a word by its collocates (Louw, 1993: 157) as well as the speaker's attitude (e.g., *naked eye*).

However, research at Birmingham has also indicated that these features of 'meaning units' are probabilistic. In the case of semantic prosody, for example, the verb 'afford' is only 'more-often-than-the-average negative' (Hunston 2011a: 81), although it occurs equally frequently in positive and negative clauses. This is largely because the overall ratio of positive to negative clauses is about 9:1 (Halliday, 1993; Matthiessen, 2006); thus, a lexical item that occurs evenly in the negative and positive is highly skewed toward the negative.

3.3 Current trends in corpus linguistics at the University of Birmingham

Much of the research in corpus linguistics carried out at the University of Birmingham has had a direct impact on language teaching. In fact, some of the researchers at Birmingham would not even call themselves 'corpus linguists,' preferring instead the more general term 'applied linguist.' This trend to pursue 'applied corpus linguistics' can be seen most easily by reviewing some of the principal works of past and present researchers at the institution, as shown below:

> Susan Hunston (Pattern Grammar, Discourse Analysis), Wolfgang Teubert (Critical Discourse Analysis, Lexicology), Geoff Barnbrook

(Local Grammar), Nick Groom (Academic Discourse, EAP), Suganthi John (Academic Discourse, EAP), Oliver Mason (Computer Linguistics), Alison Sealey (Social Linguistics, First Language Acquisition), Rosamund Moon and Gill Francis (Lexicography), Paul Thompson (Academic Discourse, EAP), Caroline Tagg (Text Analysis), Crayton Walker (Collocation), Martin Hewings (Pedagogical Grammar), Philip King (Translation Studies), and David Willis (Lexical Syllabus)

Analyzing the above research closer, it can be seen that current researchers at Birmingham can be roughly placed in one of three areas. Susan Hunston, Wolfgang Teubert, Caroline Tagg, Paul Thompson, Nicholas Groom, and Suganthi John are interested in applying corpus linguistic methodologies to discourse analysis often from an epistemological or educational viewpoint. However, although traditional approaches to corpus linguistics and discourse analysis have been seen as independent activities, the researchers at Birmingham have attempted to combine the approaches to gain a deeper understanding of texts from both a bottom-up and top-down perspective, leading also to a greater understanding of the link between language and cultural values.

Susan Hunston, Rosamund Moon, and Gill Francis have looked at applications of corpora in lexicographical works, scrutinizing the meaning and function of fixed expressions and idioms, including metaphors. In particular, a strong theme in current research is the theory of pattern grammar as illustrated in Hunston (2006), Hunston and Francis (1999), and Hunston and Sinclair (2000). Also of recent interest is the investigation of linguistic devices that express evaluation (Hunston, 2011a; Hunston & Thompson, 2000), such as nouns (e.g., *success*), verbs (e.g., *fail*), adjectives (e.g., *excellent*), adverbs (e.g., *unfortunately*), lexical bundles (e.g., *no doubt, in fact,* and *according to*), and perhaps most interestingly, lexico-grammatical patterns such as '*it was* [adjective] (e.g., *nice, kind, good, selfish, foolish*) *of* [person] *to* [do something]' (for details, see Hunston, 2011a). Importantly, Hunston (2011a: 24) shows that evaluation is often an implicit action that performed in discourse with a meaning that is contextually determined. In other words, the corpus approach only assists in the identification of evaluative meaning by pinpointing the recurrence of a particular lexical item; human interpretation is always necessary for its comprehensive analysis.

A third area is that of corpus tools and resources building, as exemplified by the work of Oliver Mason on the QTag[8] part-of-speech tagger (see also Mason, 2000; Mason & Hunston, 2004), and the creation of the Corpus Hub at Birmingham[9] (CHAB), as shown in Fig. 1. In particular, CHAB is a collection of software tools

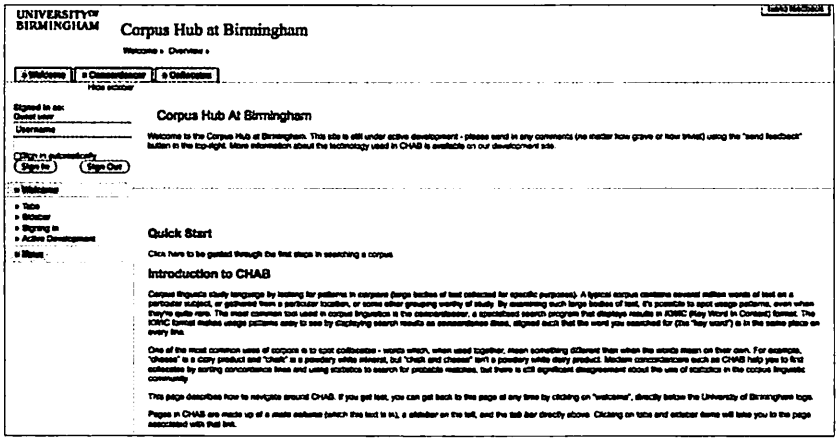

Figure 1. Corpus Hub at the University of Birmingham (CHAB)

made available under open source license, including a web-based concordancer called the Birmingham Concordancer[10], a search engine that powers the concordancer, a collection of corpora that supply the search engine[11], and a set of tools that compile the corpora, i.e., the SCAn Toolkit[12].

One recent development at the University of Birmingham has been the establishment of a dedicated center for corpus research. The Center for Corpus Linguistics (CCL) was formed in 2001, and this evolved into the current Centre for Corpus Research (CCR) in 2004. The mission of the CCR is to promote the use of corpus analysis in research, teaching, and learning[13]. Today, the CCR offers language resources (e.g., the Bank of English), facilities (e.g., a computer suite with various computer software), and technical advice to people interested in using corpora in textual studies. The CCR also hosts training workshops, seminars, and conferences on corpus research and applications in teaching. The CCR has been engaged in various projects, such as the COBUILD project with HarperCollins, lexicographical projects involving the Bank of English, academic corpus projects utilizing the British Academic Spoken English (BASE) and British Academic Written English (BAWE), and projects focused on the lexical grammar of English and corpus approaches to the study of evaluation.

3.4 The future of corpus research at the University of Birmingham

Past and present corpus studies at the University of Birmingham have prioritized the identification of meaningful language phenomena (e.g., 'meaning units') from

authentic texts. In other words, the purpose of such research in the University of Birmingham style is not always quantitative in nature. In particular, the current corpus-based studies at the University of Birmingham demonstrate how corpus linguistics and discourse analysis harmonize with one another and how this combination is an effective approach for identifying the links between language and culture. For instance, Nishina (2010) attempted to reveal the detailed cultural values of a particular discipline based on corpus-assisted discourse analysis.

A further trend in Birmingham research is to tackle the interface among corpus linguistics, phraseology, and evaluative meanings, as exemplified in Susan Hunston's recent publications (e.g., Hunston, 2011a). The direction of future research is likely to continue in this direction, with greater attention paid to extracting qualitative discourse features from corpora. In addition, the following areas are considered important: evaluation; semantic categories of lexico-grammatical patterns; links between language and (academic) culture; discourse features in sub-genres and sub-disciplines; and new word classes in English. For instance, Hunston (2011a) highlights the importance of a new class of modal-like expressions in English, based on native English speakers' preferred use of modal-like expressions over modal verbs. She also shows that learners of English consistently use modal auxiliaries more frequently than native speakers of English (Aijmer, 2002). Since modal-like expressions express evaluative meanings, by listing such expressions, one set of texts can be measured against another to compare the amount and type of such evaluative language in each. Therefore, it would be a qualitative and meaningful study to identify such evaluative features from corpora and scrutinize their discourse features.

In short, the primary essence in Birmingham school — namely, words and sequences of words are always contextual — still lives in much of the current studies at the University of Birmingham. This tradition will likely continue long into the future.

4. Corpus Linguistics at Lancaster University

4.1 The birth of corpus linguistics at Lancaster University

Corpus linguistics at Lancaster University began to grow with the creation of the Department of Linguistics and English Language in 1974. At this time the celebrated scholar Geoffrey Leech, who worked on the earlier *English Lexical Studies* project with John Sinclair, took a central role as a professor. By 1995, many more corpus linguistic researchers were working in the department, notably including Tony McEnery and Andrew Wilson. McEnery and Wilson's book *Corpus*

Linguistics (1996) is now well established as one of the most influential books in the field. In the 1990s, under McEnery and Wilson's general editorship, a series of technical papers was published by the department; McEnery and Daille (1993); McEnery and Wilson (1993); Wilson (1993); Daille (1995); Hardt-Mautner (1995); Baker (1995); Botley et al. (1996); and Takahashi (1997). This work marked the beginning of an epoch of study dealing with corpora and computational linguistics. However, there was surprisingly little overlap between the work of the corpus linguists at Lancaster and that of other researchers in the department. Today, in the same department, the situation has changed remarkably, and corpora and corpus techniques are being used by the majority of the academics to various degrees.

Another important development that has greatly shaped the corpus linguistics work at Lancaster was the creation of the UCREL research group in 1970. UCREL was initiated in the then-named Department of English by Geoffrey Leech under the name of CAMET (Computer Archive of Modern English Texts). CAMET's mission was to create a one-million-word corpus of English that would serve as a parallel British English equivalent to the Brown Corpus. With the later assistance of Universities of Oslo and Bergen, this corpus became the well-known Lancaster/Oslo-Bergen (or LOB) corpus (Johansson, Leech and Goodluck, 1978). After completing the initial LOB corpus, the researchers in the English department began collaborating with Roger Garside of the Computer Studies Department at Lancaster with the aim of developing a program to tag the corpus. This work culminated in the first version of the tagging software CLAWS[14] (Constituent Likelihood Automatic Word-tagging System), which still plays an important role in much of UCREL's research. Perhaps more importantly, the project also led to a formal agreement between the now-named Department of Linguistics and Modern English Language and the Department of Computing in 1984 to establish the Unit for Computer Research on the English Language (UCREL). The current director of UCREL, Paul Rayson, has commented that this close relationship between English language research and computer science researchers is one of the most important reasons for the success of corpus linguistics at Lancaster[15]. In fact, to reflect the growing importance of corpus-based research, in 1995, UCREL was renamed as the University Centre for Computer Corpus Research on Language, although the acronym was retained.

4.2 Core research projects at Lancaster University

Today, the Department of Linguistics and Modern English Language hosts five research clusters; theoretical and corpus-based linguistics, discourse

analysis/pragmatics and stylistics, language teaching/learning and assessment, sociolinguistics, and literacy studies. Of these, all but literacy studies involves work with UCREL. Researchers also employ a host of in-house built corpus tools and related resources, including BNCweb, BNC Web Index, CLAWS, CQPweb, LL calculator, Semantic tagger, Sentrick, and Wmatrix[16].

One of the most important projects that Lancaster researchers have been involved in is the creation of the British National Corpus (BNC). The BNC was completed in 1994 and initially released for European researchers in 1995. UCREL was a leading partner in the BNC consortium, first providing a linguistics analysis of the corpus and later providing an automated part-of-speech analysis of the entire corpus followed by a hand-corrected analysis of a two-million-word sample of the data with the aim of identifying and correcting errors and ambiguities. In 2001, the world edition of the BNC became available in Japan, followed by the XML Edition in 2007. Together with these releases, researchers at Lancaster have worked to create easy-to-use but also powerful interfaces to the corpus. The most well-known of these is the BNCweb (Hoffmann, Evert, Smith, Lee, and Berglund, 2008), which offers the powerful CQP query language from the IMS Corpus Workbench[17]. This is one of the most sophisticated tools for dealing with the BNC. Recently, Andrew Hardie has developed CQPweb, which not only replicates the functionality of BNCweb, but is also designed to work with any corpus. CQPweb is especially designed for large, fully annotated corpora that include both part-of-speech tagging and meta data. However, at present it is unable to handle corpora tagged using an XML scheme.

Reviewing the research carried out by corpus linguists at Lancaster, it is clear that much of their work is characterized by quantitative methodologies. The work of Takahashi (2006, 2007) provides a good example. In order to substantiate his claim concerning 'habitual style,' salient monologue subcorpora of the BNC (Table 1) together with the business English subcorpora of the BNC (Table 2), were examined for relevant lexical and grammatical features using multivariate techniques. In this study, it was found that individual sub-corpora showed unique features of lexical cohesion. For example, many salient words appeared in contexts associated with politics, such as *workers, patients, jobs, conditions, women, issues, members,* and *rights.* However, one reason for this was that a number of texts associated with trade union talks overwhelmed the other contexts. In addition, the overall corpus frequency of a feature may be misleading because the frequency may be the result of a few speakers using the feature very often. Therefore, Takahashi reminds corpus researchers that we have to take account of factors that may distort the frequencies of certain linguistic features, and we should establish the proportion

of corpus speakers who use the feature (regardless of frequency). We should also note such tendencies in the discussion of lexical and grammatical features, especially when we construct a Do-It-Yourself corpus.

In addition to quantitative studies, there has also been a growing trend at Lancaster University to combine quantitative and qualitative methods. For example, Paul Baker has written extensively on the results of studies that combine critical discourse analysis and corpus-linguistic approaches (e.g., Baker, 2006; 2009; 2010). Similarly, Gabrielatos, Baker, and McEnery (2010) have looked at the representation of Islam and Muslims in UK newspaper articles, using a corpus of 1,430,000 words taken from 12 national UK newspapers and their Sunday editions between 1998 and 2009. Studies of this sort can also be helped through the Wmatrix tool, developed by Paul Rayson. This tool not only allows comparison among corpora at the word level and Part-Of-Speech (POS) level, but also at the semantic level.

Table 1. Context-governed part of the BNC

Classification	Codes	Texts	Contexts
Educational and informative	scgdom1	169	Lectures, Talks, Educational demonstrations, New commentaries, Classroom interaction
Business	scgdom2	131	Company talks and interviews, Trade union talks, Sales demonstrations, Business meetings, Consultations
Public or institutional	scgdom3	262	Political speeches Sermons, Public, Government talks, Council meetings, Religious meetings, Parliamentary proceedings, Legal proceedings
Leisure	scgdom4	195	Speeches, Sports commentaries, Talks to clubs, Broadcast chat shows and phone-ins

Table 2. Demographic respondent: Monologue and Dialogue

Interaction type	codes	texts
Monologue	spolog1	212
Dialogue	spolog2	698

4.3 Current and future corpus research at Lancaster University

Lancaster University corpus researchers have been closely associated with the development of the BNC and the LOB family of corpora, as well as related tools such as CLAWS, BNCweb, and the more recent CQPweb. Research in these areas is ongoing and will no doubt continue into the future. For example, an increasing number of projects are looking at language variation over time. For these studies, the various LOB corpora have been used as a kind of 'historical' or diachronic corpus, as illustrated in the work of Baker (2009). Unfortunately, it seems unlikely that the BNC will be developed in a similar way. Today, the BNC is nearly twenty years old, and inevitably, it will become out of date. Ideally, an equivalent BNC would be built for the 2000s or 2010s, and thus combined with the current BNC we would have new large-scale diachronic corpus of British English. However, in the current economic climate, the funding for such a project seems highly unlikely.

On the other hand, Lancaster researchers are also involved in many other wide-ranging projects. Researchers such as Tony McEnery, Andrew Hardie, and Paul Rayson, for example, have had a strong interest in the applications of corpus tools and resources in the analysis of languages other than English. This is reflected in the list of current 'major' projects at UCREL that include a contrastive study of English and Chinese, the Nepali Language Resources and Localization for Education and Communication (NeLRaLEC) project, and the development of an online conceptual database of the Latin Vulgate Bible. Paul Baker and others are interested in the applications of corpus linguistics in the area of sociolinguistics, such as the *Isis: Protecting children in online social networks* project. There is also a continuing effort with UCREL to develop new tools, such as the Corpus-based grammar in contrast (CORGRAM) tool. In short, it is clear that the core essence of research in the Lancaster School, i.e., a combination of research efforts of linguists and computer scientists is still strong and will continue to flourish.

5. Corpus Linguistics at the University of Nottingham

5.1 The birth of corpus linguistics at the University of Nottingham

The birth of corpus linguistics at the University of Nottingham can be traced back to the pioneering work of Ronald Carter and Michael McCarthy. Interestingly, the early careers of both researchers were guided by the mentorship of John Sinclair at the University of Birmingham. Ronald Carter received both his MA and PhD from Birmingham, and then continued to work closely with Sinclair, serving as the prime motivator for Sinclair's book *Corpus, Concordance, and Collocation* (Sinclair, 1991) and also the editor of *Trust the Text* (Sinclair, 2004). Michael McCarthy,

who served as a lecturer at Birmingham from 1982, was closely involved in the early development of the COBUILD project. He later describes the experience as follows,

> *I became a (rather junior) colleague to an inspiring and fantastic group of people headed by Professor John Sinclair, who, to this day, is the most brilliant linguist I have ever encountered. Associated with him were legendary names such as Malcolm Coulthard, Michael Hoey, and David Brazil, and it was where I met my writing partner of so many years, Ron Carter.* (McCarthy, 2005)

Both Carter and McCarthy had a strong interest in spoken corpora and also the applications of corpora in real-world situations, in particular the classroom. Once they moved to the University of Nottingham, this led them to start developing the Cambridge and Nottingham Corpus of Discourse in English (CANCODE corpus) in the early 1990s in collaboration with Cambridge University Press. CANCODE was unique in that it was composed of spontaneous spoken discourse (in contrast to much of the spoken discourse in the British National Corpus), collected from various settings in British life, including casual conversations, people socializing together, shopping, finding out information, and general discussions[18]. With the backing of Cambridge University Press, Nottingham researchers were able to analyze the CANCODE corpus and produce many new in-class teaching materials, including *Touchstone* (McCarthy et al. 2005), *English Vocabulary in Use* (McCarthy and Dell, 1999), and *Exploring Spoken English* (Carter & McCarthy, 1997), as well as a wealth of new reference and research materials, such as *Issues in Applied Linguistics, From Corpus to Classroom* (McCarthy, 2001), and the Cambridge Grammar of English (Carter & McCarthy, 2006). The collaboration also led to the creation of the Cambridge and Nottingham Spoken Business English Corpus (CANBEC), to be discussed in more detail below, the Cambridge and Nottingham Electronic Language Corpus (CANELC), and the more recent Cambridge and Nottingham Vocational English Corpus (CANVEC).

5.2 Core research projects at the University of Nottingham

Today, corpus research at Nottingham is carried out in the Center for Research in Applied Linguistics (CRAL), which is based in the School of English Studies. Spoken discourse is still a major focus of many research projects, but the Center has expanded its interests with the creation of five different research groups, i.e., the Health Language Research Group (HLRG), the Bilingualism Research Group, Language in Professions, the Literary Linguistics Research Studio, and the Vocabulary Research Group. These broad interests are reflected in the corpora that have been developed. In addition to CANCODE, CANBEC, CANELC, and

CANVEC, two of the most important are the Nottingham Health Communication Corpus (NHCC), and the Understanding New Digital Records for eSocial Science (DReSS) corpus.

In the area of corpus linguistics, one of the key interests of researchers at Nottingham has been in the area of lexicogrammatical patterns found in spoken language, and the differences in lexicogrammatical pattern usage between spoken and written language (e.g., Carter & McCarthy, 1995; McCarthy, 1998; O'Keeffe et al., 2007). For example, 'so' is commonly used in in technical writing to signal the reason for an action, as in the sentence, 'A permanent link for the job is also provided so that the users can retrieve the result.' However, results from the CANBEC corpus reveal that 'so' is often used in spoken language as a discourse signal in a three-part turn structure of the form *head-body-tail*. For example, in the sentence, 'So, that's about hundred and fifty a month, isn't it?', 'So' serves as the *head*, and signals that the *body* of the turn contains a summary, i.e., 'that's about hundred and fifty a month.' The turn is then terminated with the *tail*, 'isn't it?' (Handford, 2010: 112). In fact, over one quarter of all occurrences of 'so' in the CANBEC corpus appear at the beginning of turns. Differences can also be found in many other aspects of language usage, such as the collocates of words, the chunks and idioms in which words are commonly found, the syntactic and semantic patterns that restrict a word's usage, and the semantic prosody (Louw, 1993) surrounding a word.

In addition to lexicogrammatical patterns of spoken language, researchers at Nottingham have investigated the range and features of spoken genres (McCarthy, 1998), the creative ways in which spoken language is used in everyday speech (Carter, 2004), and also the ways in which people communicate using non-verbal signals, such as head nods and hand gestures in lectures (Knight et al., 2009). One unique aspect of this work is a focus on the language of particular discourse communities outside the realm of academia that have traditionally been seen as difficult to access due to various issues of privacy. One example is the creation and analysis of a one-million-word corpus of emails sent by adolescents to the health website Teenage Health Freak[19], which is operated by UK-based doctors that specialize in adolescent health. After overcoming the difficulties needed to obtain permission from the website operators to conduct the study, an analysis revealed many aspects of adolescent language that had previously gone unnoticed. For example, the adolescents were shown to use many verbs commonly associated with seeking advice in a face-to-face interactions. Also, many of the adjectives used by the adolescents, such as 'afraid,' 'scared,' 'worried,' 'embarrassed,' and 'stressed.' displayed a negative perspective on the subject matter of health.

Following the general theme of Nottingham research, the results of this study have been directly applied in the training of professionals working in the health care service (Harvey et al., 2008).

Another example is the creation and analysis of the CANBEC corpus of business English, mentioned earlier. The CANBEC project began in 2001 with the aim of creating a one-million-word corpus of business English comprised of in-house and external meetings in a wide range of industries across various companies and countries. Funding was provided by the University in collaboration with Cambridge University Press, and at the completion of the corpus-building phrase in 2003, data from 65 meetings had been collected from over 20 companies, although the majority were based in the UK (Handford, 2010).

One of the major difficulties in creating the corpus was obtaining the necessary permissions from the various participants at the meetings, especially considering the sensitive nature of many of the discussions. Another difficulty was ensuring that the resulting corpus provided enough contextual knowledge, which Charles (1996) notes is an essential requirement for understanding professional discourse. To address this final issue, the corpus compiler served as a so-called 'mediating ethnographic specialist informant' (Flowerdew, 2005). This involved not only transcribing the audio and video recordings of the meetings, but also carrying out interviews and discussions with the professionals involved, collecting questionnaires to establish the context in which the meetings were being conducted, and then adding meta data to the corpus to represent these observations.

Results from the CANBEC study highlight the unique ways in which language is used in spoken business contexts. For example, Table 3 shows the key business words used in internal and external meetings with CANCODE serving as a reference corpus of general spoken discourse (Handford, 2010: 94-117). One

Table 3. Key business words in internal and external meetings in the CANBEC Corpus

1	WE	11	THOUSAND	21	BUSINESS
2	OKAY	12	HUNDRED	22	IS
3	WE'RE	13	ORDERS	23	MONTH
4	HMM	14	IF	24	STOCK
5	THE	15	WHICH	25	ISSUE
6	CUSTOMER	16	WILL	26	PRODUCT
7	NEED	17	CUSTOMERS	27	FOLLOWING
8	ORDER	18	PER	28	CENT
9	MEETING	19	PRICE	29	PROBLEM
10	SALES	20	MAIL	30	SO

Table 4. Most frequent three-word clusters in the CANBEC Corpus

Rank	Word	Rank	Word
1	I DON'T KNOW	11	IN TERMS OF
2	A LOT OF	12	ONE OF THE
3	AT THE MOMENT	13	TO DO IT
4	WE NEED TO	14	AT THE END
5	I DON'T THINK	15	I THINK IT'S
6	THE END OF	16	WE HAVE TO
7	I MEAN I	17	END OF THE
8	A BIT OF	18	I THINK WE
9	AND I THINK	19	YOU KNOW THE
10	BE ABLE TO	20	HAVE A LOOK

interesting feature of the list is the high ranking of 'we' 'we're', and 'okay.' On first sight, this may appear to highlight the collaborative nature of the business interactions. However, the addition of contextual information enabled the researchers to understand that many uses of 'we' in external meetings (i.e., those where two or more companies' representatives were involved) were in situations where the representative of a company was discussing his or her own company position and excluding the representatives of the other companies (Handford, 2010: 108-109).

Another interesting feature of business meeting interactions is the high occurrence of phrases that signal 'speculating,' 'hedging,' 'being vague,' 'specifying,' 'describing change and flux,' 'referring to collective goals,' 'protecting face,' and 'giving directives'. Some examples of these features can be seen in Table 4, which shows the most frequent three-word clusters in the corpus (Handford, 2010: 126). Comparing these results with those of academic and everyday speech, it appears that business English is similar to academic English in that it is goal driven and involves much speculating and hypothesizing. However, it is also similar to everyday speech in that it shows many signals of relationship building (Handford, 2010: 123). As above, this research has clear applications in the teaching of business practitioners and students who are preparing to enter the business world. With this in mind, Handford, M., Lisboa, M., Koester, A., and Pitt, A. (2011) have published a new textbook on business English, based on the results of the study.

5.3 The future of corpus research at the University of Nottingham

The University of Nottingham has been strongly associated with the creation

and analysis of spoken corpora in a wide range of disciples. No doubt, this trend will continue long into the future. Indeed, in an interview where Michael McCarthy was asked what he would want to spend the rest of his career working on if he had a bottomless bag of cash and unlimited personnel resources, he replied,

> *...if I had bottomless funds, surprise-surprise, I'd go on building spoken corpora (which are very costly to collect and transcribe) from as many places as I could around the world...* (McCarthy, 2005)

However, researchers at Nottingham are also become increasingly interested in the creation and analysis of multimodal corpora that represent complex interactions of verbal and non-verbal communication techniques. Creating such corpora presents an obvious challenge, but the analysis of such corpora is also difficult because most of the traditional corpus tools are designed for textual data. To address this problem, Nottingham researchers have now started developing new software tools that allow video data to be dynamically linked to traditional visualizations of transcript data, such as KWIC concordance lines.

In the same vein, researchers in computer science, engineering, and the arts and sciences have joined forces in the University's newly formed Pervasive Media Group to investigate how 'pervasive media,' such the Internet websites, social networks, and mobile phone applications, are designed, produced, and ultimately experienced by members of society. As an example of this work, Ronald Carter and other members of CRAL are developing a software toolkit that supports stand-alone and crowd-sourcing activities that will hopefully lead to new insights into contextually-appropriate uses of spoken English for non-native speakers of the language.

Nottingham's corpus researchers have always tried to ensure that their results have direct applications in language teaching, especially for non-native speakers of English. In recent years, a growing influx of foreign students into Nottingham and other European institutions, especially from China, has made this endeavor ever more important. As a result, the University is now beginning to develop ELT resources that will be delivered via web-based or mobile learning environments. A corpus analysis of these materials, as well as the language generated by learners exposed to these materials, will be a key to the success of this project.

6. Discussion

In the above sections, we reviewed the past, present, and likely future research trends at the University of Birmingham, Lancaster University, and the University of Nottingham. Comparing the corpus work at the three institutions, it can be

seen that there are many overlaps but also substantial differences. In terms of similarities, all three institutions show a strong interest in lexicogrammatical pattern usage, with Birmingham and Lancaster focusing mainly on written texts and Nottingham researchers focusing almost completely on spoken discourse. Also, there is a clear desire by researchers at all three institutions to produce results from corpus studies that lead directly to practical applications, such as the creation of language teaching materials such as dictionaries and textbooks.

When we observe the differences in the research styles and interests of the three institutions, in many ways it seems that Birmingham and Lancaster lie at opposite ends of the scale, with Nottingham lying somewhere in the middle. For example, Birmingham has traditionally focused on the analysis of very large-scale general corpora. Sinclair, in particular, has written at length on this topic and has strongly criticized alternative approaches using smaller corpora,

> *There is no virtue in being small. Small is not beautiful; it is simply a limitation. If within the dimensions of a small corpus, using corpus techniques, you can get the results that you wish to get, then your methodology is above reproach - but the results will be extremely limited...* (Sinclair, 2004: 189)

Lancaster researchers, on the other hand, have tended to see size as less crucial and have often worked with smaller corpora that are carefully *sampled* to be *representative* of the target language. Their work on the Brown family of corpora and the BNC highlight this point, as does the work of McEnery and Baker, mentioned earlier. Indeed, many corpora used at Lancaster have ranged from just a few thousand words to one million words. On the other hand, the spoken corpora created at Nottingham demonstrate the usefulness of both approaches, with CANCODE serving as a very large-scale corpus of general English, and CANBEC and other corpora serving as much smaller one-million word corpora that are more representative of language features in narrow disciplines, such as business.

A difference in views has also been noted in regard to the use and importance of annotation and corpus meta data. Traditionally, this has often been expressed as a difference between 'corpus-driven' and 'corpus-based' approaches. Again, Birmingham scholars have tended to adopt a more radical view on this topic, with Sinclair of the 'corpus-driven' school of corpus linguistics expressing the following view,

> *The interspersing of tags in a language text is a perilous activity, because the text thereby loses its integrity, and no matter how careful one is the original text cannot be reliably retrieved. [In tagged corpora] The corpus data can only be observed through the tags; that is to say, anything the tags are not sensitive to will be missed...In corpus-driven linguistics you do not use pre-tagged text, but you*

process the raw text directly and then the patterns of this uncontaminated text are able to be observed. (Sinclair, 2004:191)

However, it should be noted that Susan Hunston, who is one of the current leading researchers at Birmingham's Centre for Corpus Research (CCR), has recently stated that annotated corpora are in many cases justified and useful (Hunston, 2011b).

The view of Lancaster scholars is starkly different to that of Sinclair and much more in line with that of Hunston. McEnery and Hardie (2012: 163), for example, provide a strong defense of annotation procedures while dismissing many of Sinclair's concerns about annotation and the 'corpus-based' approach as either farcical, unsupported, or in the worst case, banal. They emphasize, however, that this is an evaluation of Sinclair's stance on theoretical issues in corpus research, but should not detract in any way from his contribution to corpus linguistics in terms of new methods, insights, and discoveries. The researchers at Nottingham, again, seem to adopt the middle ground. Many of their corpora use highly complex annotation and meta-data schemes, as exemplified by the Understanding New Digital Records for eSocial Science (DReSS) corpus. Indeed, it would be impossible to carry out research of this kind without resorting to annotation schemes that 'contaminate' the corpus, using the words of Sinclair. However, they are also happy to carry out analyses of raw texts, as illustrated by the findings of Harvey et al., (2008) using the Nottingham Health Communication Corpus (NHCC).

One final difference between the three institutions is revealed by looking at their funding procedures. In this case, however, all three universities show unique traits. Much of the early corpus work at the University of Birmingham was funded through a collaboration with the publisher Harper (now HarperCollins). This close relationship allowed rapid progress to be made and resulted in the publication of multiple language resources, with the COBUILD dictionary being the most notable. However, when the Harper funding was eventually terminated, the project became far more difficult to maintain, and many researchers sadly had to leave the project. Since then, while external funding still plays a role in the development of projects, it is clearly less important than in the past. This point is highlighted by viewing the homepage of Birmingham's Centre for Corpus Research, where there is no mention of external funding grants or projects. Rather, the Centre appears to be a more independent organization, generating funding partially through organizing training workshops and carrying out consultancy work for outside clients.

Corpus researchers at Lancaster University, on the other hand, have funded most of their projects through government grants, such as those issued by the UK Arts and Humanities Research Council (AHRC), the UK Engineering and Physical

Sciences Research Council (EPSRC), and the European Union (EU). Again, the advantage of this strategy is that they can be less dependent on commercial pressures. However, projects tend to be shorter, and with the recent education cuts faced by many European institutions, the competition for grants has obviously increased. Fortunately, corpus researchers who are members of Lancaster University's UCREL group have more opportunities than many others working in arts and humanities, due to the interdisciplinary nature of their work.

Corpus research at the University of Nottingham has been greatly assisted with funding from Cambridge University Press. The financial backing from such a strong commercial institution as CUP has clearly benefitted the Nottingham researchers greatly, and the resulting corpora, language resources, and textbooks are impressive. To what extent this collaboration will continue into the future is an unknown, but the list of current and new projects initiated with financial backing from CUP that are listed on the Centre for Research in Applied Linguistics (CRAL) website suggests that the future is a bright one.

7. Final Thoughts

This paper emerged from a symposium of the same title held at the Japan Association of English Corpus Studies (JAECS) national conference in October 2011. During the organization of the seminar and the writing of this paper, it became increasingly clear how much the three institutions discussed in this paper, i.e., the University of Birmingham, Lancaster University, and the University of Nottingham, have influenced the work of corpus researchers in Japan. As an example, the normally relatively easy task of deciding suitable panelists to discuss the work at each institution became surprisingly difficult due to the sheer number of excellent scholars working in Japan that have graduated from one of these three institutions. Also, a simple review of the references included in past and present works published in the journal of the Japan Association of English Corpus Studies (JAECS) reveals how many Japanese scholars continue to read and build on the work of researchers at Birmingham, Lancaster, and Nottingham. Clearly, many scholars in Japan have found that analyzing plain-text corpora, the approach advocated by Sinclair and others in the Birmingham school, can be extremely insightful. Many others have capitalized on the advances in part-of-speech tagging and other meta-data processing techniques to make useful discoveries about language in a similar way to researchers from the Lancaster school. Others again, have started taking a serious look at spoken discourse, clearly influenced by the major contributions made by researchers at Nottingham. It should be stressed, however, that the

differences in approaches across the three schools mentioned above are only abstractions, and in fact, many researchers at all three institutions will have looked at both written and spoken discourse using a combination of small and large corpora that are either plain text of annotated.

In this paper, we have attempted to present a balanced review of the work at three British research institutions that have made significant contributions to the field of corpus linguistics. Of course, these are not the only British institutions to have contributed to the field. The University of Edinburgh, the University of Strathclyde, the University of Liverpool, Swansea University, Oxford Brookes University, the University of Reading, University College London, University of Warwick, the University of Leeds and many others have also made important discoveries. However, we hope that this review of the work at Birmingham, Lancaster, and Nottingham can provide some insights into how the field developed and flourished in Britain in the 1960s, 1970s, and 1980s, became a mainstream research discipline in its own right in the 1990s, and advanced further in the 2000s. There is no doubt that today we are experiencing a 'golden age' of corpus linguistics, and the contributions of researchers at Birmingham, Lancaster, and Nottingham research cannot be understated.

Acknowledgements

We would like to express our appreciation to the many members of the faculty at the three institutions represented in this study for their advice, suggestions, and clarifications of certain issues. In particular, we would like to thank Tony McEnery and Paul Rayson of Lancaster University, and Susan Hunston of the University of Birmingham for their generous offers of help during the research for this paper.

Notes

1. http://icame.uib.no/brown/bcm.html
2. http://www.titania.bham.ac.uk/
3. http://www.natcorp.ox.ac.uk/
4. http://khnt.hit.uib.no/icame/manuals/londlund/index.htm
5. http://www.lexically.net/software/index.htm
6. http://www.lexically.net/wordsmith/index.html
7. http://bncweb.info/
8. http://morphix-nlp.berlios.de/manual/node17.html
9. For more information on CHAB, visit: https://arts-ccr-002.bham.ac.uk/chab/
10. The Birmingham Concordancer is a web application that offers various tools

from modern concordancers.

11. A large collection of tools for compiling corpora has been developed as part of CHAB, particularly for use with SCAn.
12. The SQL-Based Corpus Analyser (SCAn) is a corpus search engine that stores corpora in a MySQL database.
13. The current director of the CCR is Professor Paul Thompson; the technical director is Dr. Oliver Mason.
14. http://ucrel.lancs.ac.uk/claws/
15. Personal communication.
16. http://ucrel.lancs.ac.uk/wmatrix/
17. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/index.html
18. http://www.cambridge.org/de/elt/catalogue/subject/custom/item3646595/Cambridge-International-Corpus-Cambridge-and-Nottingham-Corpus-of-Discourse-in-English-%28CANCODE%29/?site_locale=de_DE
19. http://www.teenagehealthfreak.org

References

Aijmer, K. (2002) "Modality in Advanced Swedish Learners' Written Interlanguage." In S. Granger, J. Hung, & S. Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Amsterdam: Benjamins, pp.55-76.

Anthony, L. (2011) AntConc (Version 3.2.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Baker, P. (1995) *The Evaluation of Multiple Post-Editors: Inter-Rater Consistency in Correcting Automatically Tagged Data* (Technical Report of University centre for computing corpus research on language), Lancaster University.

---------- (2006) *Using Corpora in Discourse Analysis.* London: Continuum.

---------- (2009) "The BE06 Corpus of British English and Recent Language Change." *International Journal of Corpus Linguistics*14, 3:312-337.

---------- (2010) *Sociolinguistics and Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Botley, S., Glass J., McEnery T., & Wilson, A. (1996) Special Issue. Proceedings of Teaching and Language Corpora 1996. (TALC96) [Online]. URL:http://ucrel.lancs.ac.uk/talc96/

Carter, R. (2004) *Language and Creativity: the Art of Common Talk.* London: Routledge.

Carter, R. & McCarthy, M. (1995) "Grammar and the Spoken Language." *Applied Linguistics*16, 2:141-158.

---------- (1997) *Exploring Spoken English.* Cambridge: Cambridge University Press.

---------- (2006) *Cambridge Grammar of English.* Cambridge: Cambridge University Press.

Charles, M. (1996) "Business Negotiations: Interdependence between Discourse and the Business Relationship," *English for Specific Purposes*15, 19-36.

---------- (2004) *The Construction of Stance: A Corpus-Based Investigation of Two Contrasting Disciplines.* Unpublished doctoral dissertation, University of Birmingham, Birmingham, UK.

---------- (2006a) "Phraseological Patterns in Reporting Clauses Used in Citation: A Corpus-Based Study of Theses in Two Disciplines." *English for Specific Purposes*25, 3:310-331.

---------- (2006b) "The Construction of Stance in Reporting Clauses: A Cross-Disciplinary Study of Theses." *Applied Linguistics*27, 492-518.

Erman, B., & Warren, B. (2000) "The Idiom Principle and the Open Choice Principle." *Text*20, 1:29-62.

Firth. J. R. (1935) "The Technique of Semantics." *Transactions of the Philological Society*7, 36-72.

---------- (1957) *Papers in Linguistics 1934-1951.* Oxford: Oxford University Press.

Flowerdew, L. (2005) "An Integration of Corpus-Based and Genre-Based Approaches to Text Analysis in EAP/ESP: Countering Criticisms against Corpus-Based Methodologies." *English for Specific Purposes*24, 321-332.

Gabrielatos, C., Baker, P., & McEnery, T. (2010) Using Sketch Engine to Examine the Presentation of Islam and Muslims in the UK press. 43rd Annual Meeting of the British Association for Applied Linguistics (BAAL), 9-11 September 2010, University of Aberdeen. [Online].
URL:http://eprints.lancs.ac.uk/34198/

Halliday, M.A.K. (1993) "Quantitative Studies and Probabilities in Grammar." In M. Hoey (ed.), *Data description discourse.* London: HarperCollins, pp. 1-25.

Handford, M. (2010) *The Language of Business Meetings.* Cambridge: Cambridge University Press.

Handford, M., Lisboa, M., Koester, A., & Pitt, A. (2011) *Business Advantage.* Cambridge: Cambridge University Press.

Hardt-Mautner, G. (1995) "Only Connect. Critical Discourse Analysis and Corpus Linguistics.", *Technical Report of University Centre for Computing Corpus Research on Language,* Lancaster University.

Harvey, K., Churchill, D., Crawford P., Brown B., Mullanya, L., Macfarlane A., & McPherson A. (2008) "Health Communication and Adolescents: What Do

Their Emails Tell Us?" *Family Practice*25, 4:304-311.

Hewings, M. (2011, October 21-23) *Using Corpora in Research, Teaching, and Materials Design for ESP: An Evaluation.* Unpublished Manuscript, Keynote Address Presented at the 2011 International Conference on English for Specific Purposes (ICESP 2011) . HungKuang University, Taichung, Taiwan.

Hoffmann, S., Evert S., Smith N., Lee D., & Berglund Y. (2008) *Corpus Linguistics with BNCweb: A Practical Guide.* Peter Lang AG.

Hunston, S. (2006) "Phraseology and System: A Contribution to the Debate." In G. Thompson & S. Hunston (eds.), *System and Corpus: Exploring Connections* London: Equinox, pp.55-80.

---------- (2011a) *Corpus Approach to Evaluation: Phraseology and Evaluative Language.* New York: Routledge.

---------- (2011b, November 26) *Corpus Approaches to the Study of Evaluation.* Unpublished Manuscript, Special Seminar of the Japan Association of English Corpus Studies (JAECS). November 26, 2011. Kyoto, Japan.

Hunston, S., & Francis, G. (1999) *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English.* Amsterdam: John Benjamins.

Hunston, S., & Sinclair, J. (2000) "A Local Grammar of Evaluation." In S. Hunston & G. Thompson (eds.), *Evaluation in Text: Authorial Stance and the Construction of Discourse.* Oxford: Oxford University Press, pp.74-101.

Hunston, S., & Thompson, G. (2000) *Evaluation in Text: Authorial Stance and the Construction of Discourse.* Oxford: Oxford University Press.

Johansson, S., Leech, G., & Goodluck, H. (1978) *Manual of Information to Accompany the Lancaster-Olso/Bergen Corpus of British English, for Use with Digital Computers.* Department of English, University of Oslo.

John, T. (1986) "Micro-Concord: A language Learner's Research Tool." *System*14, 151-162.

Knight, D., Evans, D., Carter, R., & Adolphs, S. (2009) "HeadTalk, HandTalk and the Corpus: Towards a Framework for Multi-Modal, Multi-Media Corpus Development." *Corpora Journal*4, 1:1-32.

Louw, B. (1993) "Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies." In M. Baker, G. Francis & E. Tognini-Bonelli (eds.), *Text and Technology: In Honour of John Sinclair.* Amsterdam: John Benjamins, pp.157-176.

Mason, O. (2000) *Programming for Corpus Linguistics: How to Do Text Analysis with JAVA.* Edinburgh: Edinburgh University Press.

Mason, O., & Hunston, S. (2004) "The Automatic Recognition of Verb Patterns: A Feasibility Study." *International Journal of Corpus Linguistics*9, 2:253-270.

Matthiessen, C. (2006) "Frequency Profiles of Some Basic Grammatical Systems: An Interim Report." In G. Thompson & S. Hunston (eds.), *System and Corpus: Exploring Connections.* London: Equinox, pp.103-142.

McCarthy, M. (1998) *Spoken Language and Applied Linguistics.* Cambridge: Cambridge University Press.

---------- (2001) *Issues in Applied Linguistics, From Corpus to Classroom.* Cambridge: Cambridge University Press.

---------- (2005) Interview with Michael McCarthy ELTNEWS.com. [Online]. URL:http://www.eltnews.com/features/interviews/2005/01/interview_with_michael_mccarth.html

McCarthy, M., McCarten, J., & Sandiford, H. (2005) *Touchstone.* Cambridge: Cambridge University Press.

McCarthy, M., O'Dell, F. (1999) *English Vocabulary in Use.* Cambridge: Cambridge University Press.

McEnery, T. & Daille, B. (1993) *Database Design for Corpus Storage: The ET10-63 Data Model* (Technical Report of University centre for computing corpus research on language), Lancaster University.

McEnery, T. & Hardie, A. (2011) *Corpus Linguistics.* Cambridge: Cambridge University Press.

McEnery, T. & Wilson, A. (1993) *Corpora and Translation: Uses and Future Prospects* (Technical Report of University centre for computing corpus research on language), Lancaster University.

---------- (1996) *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Nattinger, J. R., & DeCarrico, J. S. (1992) *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Nishina, Y. (2010) *Evaluative Meaning and Disciplinary Values: A Corpus-Based Study of Adjective Patterns in Research Articles in Applied Linguistics and Business Studies.* Doctoral dissertation, University of Birmingham.

Nunex, P. P. (2006) "An Interview with Geoffrey Leech." *Altantis*29, 1:143-156.

O'Keeffe, A., McCarthy, M. & Carter, R. (2007) *From Corpus to Classroom: Language Use and Language Teaching.* Cambridge: Cambridge University Press.

Palmer, H. & A. S. Hornby. (1933) *The Second Interim Report on English Collocations.* Tokyo: Kaitakusha.

Sinclair, J. M. (1991) *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

---------- (2004) *Trust the Text: Language, Corpus and Discourse.* London: Routledge.

Sinclair, J. M. (ed.) (1987) *Collins COBUILD English Language Dictionary* (1st

ed.). London: Williams Collins Sons & Co Ltd.

Sinclair, S., Jones, S., Daley, R., & Krishnamurthy, R. (2004) *English Collocation Studies: The OSTI Report.* London and New York: Continuum.

Stubbs, M. (2001) *Words and Phrases.* Oxford: Blackwell.

Takahashi, K. (2006) "A Study of Register Variation in the British National Corpus." *Literary and Linguistic Computing*16, 1:111-126.

---------- (2007) *Typology of Registers in the British National Corpus: Multi-Feature and Multi-Dimensional Analyses.* Doctoral dissertation, Lancaster University.

Teubert, W. (2003) "Writing, Hermeneutics, and Corpus Linguistics." *Logos and Language* 4, 1-17.

Wilson, A. (1993) *Towards an Integration of Content Analysis and Discourse Analysis: The Automatic Linkage of Key Relations in Text* (Technical Report of University centre for computing corpus research on language), Lancaster University.

(Waseda University: anthony@waseda.jp)
(Meiji Gakuin University: yasunori.nishina.yn@googlemail.com)
(Toyota National College of Technology: takahasi@toyota-ct.ac.jp)
(University of Tokyo: mike@civil.t.u-tokyo.ac.jp)