Variation and Change in Spoken and Written Discourse

Dialogue Studies (DS)

*Dialogue Studie*s takes the notion of dialogicity as central; it encompasses every type of language use, workaday, institutional and literary. By covering the whole range of language use, the growing field of dialogue studies comes close to pragmatics and studies in discourse or conversation. The concept of dialogicity, however, provides a clear methodological profile. The series aims to cross disciplinary boundaries and considers a genuinely inter-disciplinary approach necessary for addressing the complex phenomenon of dialogic language use. This peer reviewed series will include monographs, thematic collections of articles, and textbooks in the relevant areas.

For an overview of all books published in this series, please see
http://benjamins.com/catalog/ds

Volume 21

Variation and Change in Spoken and Written Discourse
Perspectives from corpus linguistics
Edited by Julia Bamford, Silvia Cavalieri and Giuliana Diani

Variation and Change in Spoken and Written Discourse

Perspectives from corpus linguistics

Edited by

Julia Bamford
Università di Napoli "L'Orientale"

Silvia Cavalieri
Università di Milano

Giuliana Diani
Università di Modena e Reggio Emilia

John Benjamins Publishing Company
Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Variation and Change in Spoken and Written Discourse : Perspectives from corpus linguistics / Edited by Julia Bamford, Silvia Cavalieri and Giuliana Diani.
p. cm. (Dialogue Studies, ISSN 1875-1792 ; v. 21)
Includes bibliographical references and index.
1. Discourse analysis. 2. Academic writing. 3. Interpersonal communication. 4. Linguistic change. 5. Language and languages--Variation. I. Bamford, Julia. II. Cavalieri, Silvia. III. Diani, Giuliana.
P302.V363 2013
401'.41--dc23 2013028679
ISBN 978 90 272 1038 8 (Hb ; alk. paper)
ISBN 978 90 272 7121 1 (Eb)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

Acknowledgements

The editors of this book would like to take this opportunity to thank all the people who have made this volume possible. To begin with, we would like to express our deepest thanks to our colleagues in the research group CLAVIER, who made the organisation of the CLAVIER Conference a reality.

Special thanks go to all the participants in the present volume who helped to make the Conference a fruitful forum of discussion around variation and change in language use in spoken and written discourse, and who with brilliance and hard work have shaped their valuable contributions to the Conference as chapters for the present volume.

We are also most grateful to the members of the scientific committee who evaluated the proposals for the CLAVIER Conference. Their valuable comments, suggestions and feedback have undoubtedly provided added value to the articles in this book and greatly contributed to its overall quality.

Last but not least, our thanks go to the *Dialogue Studies Series* Editor, Professor Edda Weigand, and to the publisher John Benjamins for believing in this volume and for assisting us in its completion.

Introduction

Giuliana Diani
University of Modena and Reggio Emilia, Italy

This volume contains a selection of sixteen papers from the CLAVIER Conference on "Corpus Linguistics and Language Variation", held in Modena (Italy) in November 2009. The Conference was hosted by the organizing committee of the CLAVIER research group (Corpus and Language Variation in English Research group), a research centre founded in 2009 by the Universities of Bergamo, Florence, Milan, Modena and Reggio Emilia, Rome "Sapienza", Siena and Trieste, and currently based in Modena.

The volume focuses on aspects of variation and change in language use in spoken and written discourse on the basis of corpus analyses, providing new descriptive insights, and new methods of utilising small specialized corpora for the description of language variation and change. All the contributions represent a variety of diverse views and approaches, but all share the common goal of throwing light on a crucial dimension of discourse: the dialogic interactivity between the spoken and written. The contributions selected for this book not only witness the interest in examining discourse from the point of view of its dialogic qualities using corpus methods, but also show the breadth and depth of the field. Their focuses range from papers addressing general issues related to corpus analysis of spoken dialogue to papers focusing on specific cases employing a variety of analytical tools, including qualitative and quantitative analysis of small and large corpora. Moreover, the book considers the time dimension with some contributions looking at the relationship between spoken and written discourse from a diachronic perspective.

The chapters of the book can be divided into two parts, which highlight specific aspects of corpus analysis in spoken and written discourse. The first deals with corpus analysis of spoken dialogue, with papers whose focuses range from issues related to language variation in spoken academic and business discourse to papers focusing on dialogic variation and language varieties. The second presents a number of specific case studies based on written corpora addressing language change from a diachronic perspective.

Overview of the chapters

The first two chapters of Part I ('Corpus analysis of spoken dialogue') focus on language variation in spoken academic discourse. The opening article, by Anna Mauranen, explores issues of corpus use, with particular focus on spoken corpora of academic language. More specifically, the chapter focuses on the use of English as a Lingua Franca in academic settings (ELFA). Drawing on the experience of compiling and analysis of the ELFA corpus, comprising academic speech (ELFA: www.eng.helsinki.fi/elfa), Mauranen's study tackles issues of data selection, relevance, and meaningful combinations of analytical methods. Her aim is to show that corpus methods have a lot to offer in teasing out the big picture and emergent patterning from the bewildering detail that small-scale studies easily drown themselves in. However, Mauranen suggests that they require a good database in order to yield good answers. The chapter provides evidence that it is important to focus on corpus compilation sharply so as to keep the effort tolerable while getting the most out of the data.

The second article, by Akiko Okamura, investigates how speakers employ personal pronouns (*we*, *you*, *I*) in two types of monologic academic speech, undergraduate lectures and public lectures, through the analysis of the Michigan Corpus of Academic Spoken English (MICASE). Her study demonstrates that the frequency of use of personal pronouns is greatly influenced by the type of academic speech. It also shows that both common features and variations in academic speech are due to its purpose and the relationship between the speaker and the audience. Her findings suggest that common features are related to characteristics of oral presentation, observed in the linguistic environment of the personal pronouns.

A second trend of Part I is represented by two chapters, by Janet Bowker and Belinda Crawford Camiciottoli, tracing the concept of dialogue in spoken and written business discourse. Janet Bowker explores convergences and divergences in cross-register dynamics as displayed in the language of corporate communications, and more specifically in the messaging networks of in-house, internal company interactions between management and the workforce. Her discussion of examples not only identifies how written and visually presented information conditions the spoken language of company oral presentations in relation to its communicative purposes but also how the features of spoken discourse influence the language and pragmatic impact of company e-distributed newsletters. Belinda Crawford Camiciottoli's chapter illustrates an application of grammatical tagging as a methodological tool for the investigation of small specialized spoken and written corpora: spoken earnings presentations and written earnings releases. The analysis focuses on two key features: lexical density and evaluative

adjectives. Her results reveal interesting differences between the two corpora that appeared to be influenced by mode, interactional setting, and role/status of speakers and writers. The chapter shows how grammatical tagging offers new ways to integrate quantitative and qualitative methods in order to better understand discourse used in specific communicative contexts.

A third trend is represented by four chapters addressing specific issues referring to dialogic variation and language varieties. Michelina Savino explores regional variation in Italian question intonation from a corpus perspective. She examines a section of the CLIPS corpus (Corpora e Lessici di Italiano Parlato e Scritto, *Corpora and Lexicons of Spoken and Written Italian*) consisting of a collection of Map Task dialogues of Northern, Central, and Southern accents estimated as representative of Italian regional variation. Her results show that the most widespread intonation pattern for questions is rising-falling (not falling-rising), and the distribution of the rising-falling and falling-rising contour types across varieties is not regionally conditioned.

In the next chapter, Rene Altrov and Hille Pajupuu analyse a corpus of emotional speech. Their corpus (The Estonian Emotional Speech Corpus) aims at serving as an acoustic basis for corpus-based synthesis of emotional speech from text. They exemplify each emotion by a hundred sentences with no content influence on emotion identification. From their analysis, it emerges that emotions can be identified in non-acted speech.

The contribution by Pierfranca Forchini focuses on movie corpora to explore spoken American English by applying Biber's (1988) Multi-Dimensional approach. Her study illustrates an experiment with 3rd year Italian students of English that proves the potentiality of this approach especially in the learning of elisions, blends, repetitions, false starts, reformulations, discourse markers, and interjections.

Wendy Anderson's chapter analyses the ways in which geographical variation can be explored both quantitatively and qualitatively using the Scottish Corpus of Texts & Speech (SCOTS). Her study gives an overview of the geographically-defined varieties of Scots represented in the corpus under investigation, and demonstrates how the complex web of variation can be analysed quantitatively using integrated corpus tools.

With the second part of the volume our attention is drawn to investigations of written corpora from a diachronic perspective ('Using corpora to analyse written discourse: a diachronic perspective'). The first four articles employ diachronic approaches to historical corpora. The first contribution, by Udo Fries, discusses the possibilities for research with the Zurich English Newspaper Corpus (ZEN) and ways of expanding this corpus. His study deals with a special collection of newspapers within the ZEN Corpus, the papers of January 1701. Through the

analysis of six newspapers, he identifies some aspects of variation (morphological and text-linguistic). Besides the study of grammatical variation, the analysis gives – linguistic – answers to a classification of early English newspapers.

The second article, by Gabriella Del Lungo Camiciotti, analyses Ruskin's guidebook *Mornings in Florence* with a view to investigating how heritage sites and places are construed from the writer's point of view in the context of the development of modern travel guides from diaries and personal notes to works addressing a wide audience of tourists. Her analysis suggests that the perception and textual construction of space varies in accordance with shifting cultural frameworks and world views.

Paul Tucker, in the next chapter, examines the character and function of 'justificatory arguments' in writing on visual art following Toulmin ([1958] 2003)'s model on the uses of argument. His study tests the model's applicability to aesthetic discourse by examining a small historical corpus of exhibition reviews. His analysis shows that, as prescribed by the model, claims are there supported by arguments whose relevance is underwritten by warrants, though mostly these are tacitly invoked. It also reveals synchronic and diachronic variation in the kind of warrant invoked, in apparent correspondence to a historical shift in the kind of statement prevalently used to make aesthetic claims.

The contribution by Chiara Prosperi Porta investigates a small corpus of biostatistics from the point of view of its evolution in terms of textual organisation and models. She explores the diachronic variations in the conceptual encoding of the discipline, its methodology and the grammatical structures used in the presentation, argumentation and interpretation of numerical data applied to the bio sciences. Her findings show that variation is reflected in the corpus according to the respective discourse communities and diverse communicative purposes across time.

A second trend of Part II is represented by four articles dealing with diachronic methodologies and language change. Šarolta Godnič Vičič's chapter explores discoursal change in research article abstracts in tourism studies. Based on a corpus of research article abstracts published over a span of thirty years in three prominent academic journals, she investigates changes in the patterns of use of the demonstrative 'this'. Her findings show that the demonstrative is increasingly used with a narrow range of lexical items which seem to signal change in the way authors introduce their research to the discourse community and persuade readers to continue to read the research article.

In the next chapter, Stefania Spina examines changes in the frequency and use of some selected linguistic features in the language of Italian printed news: left dislocations, sentence-initial connectives, sentence length, lexical density and subordinating conjunctions. Her study adopts a diachronic approach and relies on a corpus-based methodology. She measures language change between 1985 and

2000 using two sub-sections of the *Repubblica* corpus. Her data show that in the time-frame between 1985 and 2000 there are emergent trends of linguistic change regarding specific linguistic features.

The contribution by Tadaharu Tanomura analyses diachronic changes of the grammar and expressions of contemporary Japanese based upon the texts of the minutes of the National Diet of Japan. From the analysis, it emerges that the minutes of the National Diet of Japan is an invaluable source of information for diachronic research of contemporary Japanese. Through the analysis of texts of daily newspapers, the study also reveals periodical (e.g., seasonal, monthly and weekly) changes in language use.

Finally, Caroline Clark's diachronic study compares two large contemporary corpora of British quality newspapers by investigating the increased popularisation of newspaper register. The study focuses on those examples which are highlighted by a quantitative comparative overview of the two corpora based on a series of analyses using keyword and concordancing tools. Her results show that a shift in presentation and style is present, with an increased 'familiarisation' of language, in particular the use of spoken forms.

As illustrated in this brief overview, the analyses collected in this volume confirm that corpora represent a powerful analytical tool both in applied and theoretical linguistics. They are of particularly significant importance in studies on language variation and language varieties. The wealth and amount of data made available through corpus compilation and query tools have enabled scholars to explore differences across spoken and written discourse, diachronic and geographic varieties.

PART I

Corpus analysis of spoken dialogue

SECTION I

Variation and academic dialogue

CHAPTER 1

Speaking professionally in an L2

Issues of corpus methodology

Anna Mauranen
University of Helsinki, Finland

The fastest-growing use of globalised English is among speakers who do not share a first language, that is, English used as a lingua franca (ELF). To keep up with the developments of the language in such varying circumstances poses a challenge to research: how can we access reliable data that captures new directions in this expanding use of English? How should we go about securing enough data in a new area of language use, where variability is highly unpredictable, and change is likely to be fast? Clearly, corpus methods have a lot to offer in teasing out the big picture and emergent patterning from the bewildering detail that small-scale studies easily drown themselves in. ELF has established itself particularly in two important and influential inherently highly international domains: science and business. Both are high-stakes domains where language plays an important role. It makes sense to pay close attention to the ways English works in them and how it takes shape. This paper looks into the scientific sphere, and draws on the experience of compiling and analysis of the first ELF corpus, comprising academic speech (ELFA: www.eng.helsinki.fi/elfa). It will tackle issues of data selection, relevance, and meaningful combinations of analytical methods.

1. Introduction

Large corpora have become a mainstream tool in linguistic enquiry in the last two decades. This period overlaps roughly with the emergence of spoken language at the centre of attention in linguistic enquiry. Yet big is not always beautiful: the two have come together more rarely than one would wish, given that both have been remarkably influential in shaping contemporary perceptions of language. Both have also taken new departures from the trodden path in many domains of applied linguistics. Lexicography, translation, and language teaching have benefited enormously from corpora, and so has the teaching of special languages – but overwhelmingly in the written mode. The lively research in spoken language that

has developed in qualitative research traditions, such as discourse analysis, conversation analysis, or interactional linguistics has also enlivened language teaching, but not found its way into corpus linguistics on a large scale – despite notable exceptions like the work by scholars such as Stenström (1995), Biber (1988, 2006), Aijmer (2002), and Carter and McCarthy (2006), to name but a few pioneers.

Corpora made their way to research in professional and academic language (see e.g., Hyland 1998, 2000; Bondi 1999) on account of the perceived interests of students and academics to read and eventually publish in English. This research makes an important contribution to ESP – but again remains in the written domain. Yet even a brief glance at the multifarious environments of academic language suffices to reveal that both speaking and writing are at stake. The first spoken corpora in academic English began to get compiled in the late 1990s, MICASE (http://quod.lib.umich.edu/m/micase) and T2K-SWAL (www.ets.org/Media/Research/pdf/RM-04-03.pdf), with an original motive in the practical needs of language testing. Both were located in US universities, with a clear focus on native speakers of English, as was the case with their later British counterpart BASE (www.warwick.ac.uk/fac/soc/al/research/collect/base/). It is only very recently that the self-evident primacy of the native speaker of English has been questioned in academic and professional contexts. But since the turn of the millennium, a reconceptualisation of international English as one of the most important new departures from traditional orientation in linguistics and applied linguistics has gained ground (see Widdowson 1994; Jenkins 2000, 2007, 2013; Seidlhofer 2001, 2011; Mauranen 2006a, 2010, 2012) – and begun to compile its own corpora. The ELFA corpus of academic spoken English (www.eng.helsinki.fi/elfa/elfacorpus.htm) is the first, and so far the only, large database based on English used as a lingua franca in academic contexts.

Corpora tend to be laborious to compile, and speech corpora invariably involve an enormous amount of work before they are accessible to the research community. The same corpora are therefore normally used by a large number of researchers over considerable time, unlike smaller data samples gathered by individual scholars for their personal use. It is particularly pertinent to engage with principles of compilation and utilisation of such widely shared data.

This paper looks into the methodological repercussions of this new departure in English corpus linguistics: combining spoken corpora and English as an international lingua franca. The context is academia, one of the major sites of English as a globally influential lingua franca, and the key environment where socialisation into professions takes place.

2. Background

The analysis of professional language, with English as the overwhelmingly most widely used language, originated in the needs of teaching, as is very clear in the early studies (see e.g., Swales 1985; Trimble 1985). For a long time, research into professional English was very strongly oriented to the written mode, and as the needs of students in higher education in different countries was the target application, reading and writing English were prime concerns. The academic world was predominantly seen from the perspective of the written, mostly printed, word. Reading texts for study and writing for achieving qualifications and positions were perceived to be the topmost needs of students and novice academics who were preparing for their future profession or a university career. The fast-rising number of student mobility and exchange programmes that really gained momentum after the turn of the millennium as well as the ever-growing number of international conferences have raised awareness of the centrality of spoken skills in academia. The first corpora of spoken academic English in the late 1990s in the USA were a response to the pressing needs of testing prospective students' and teaching assistants' ability to cope with spoken interaction in an English-speaking environment.

Spoken EAP corpora in the U.S., MICASE and T2K-SWAL, thus reflected a shift in awareness in teaching and testing academic language: assessment and appropriate support to large numbers of students required research-based solutions in the domain of speaking just as much as in writing. The BASE corpus soon followed suit in the UK. The idea was, in line with the study and teaching of written skills, that observing closely what speakers with English as their native language (ENL) do would yield the best basis for teaching and assessing students who spoke other first languages. In an environment where English is the main language of the university and the community at large, this was not an unreasonable point of departure. However, a more global look at English paints a very different picture, with its varying linguistic landscapes (cf. Jenkins 2013). Not only do we live amidst proliferating international exchange and degree programmes, but the language of equally endlessly expanding conferencing is English. Investigating the language of conferences (see e.g., Ventola et al. 2002) makes an important contribution, but so far it has been based on the tacit assumption that however international academic conferences are, ENL models are the appropriate targets for language use. Yet in reality the conference language is ELF, and the "expert users" (Rampton 1990) of language are a more relevant target than speakers of a particular L1, where effectiveness is the real target (for conference ELF, see also Mauranen 2013).

The interest in spoken discourse does not stop at its practical usefulness any more than that of writing; both hold much promise for the scholar who seeks to understand discourse in academia, and speaking is at least as much a key to making sense of academic discourse as writing. It is a crucial ingredient in maintaining social structures. Academic institutions engage in constant talk: we hold lectures, seminars, and consultations as part of our pedagogical duties, we organise conferences, panel discussions and public presentations, we give talks at graduation and other ceremonies, and we talk our way through endless meetings. In talk we maintain, negotiate and reproduce our institutional relations at the various administrative levels of our organisations. We can see this working at the macro-social level as repeated action that creates and maintains social structures, much in the way described in Giddens's (1984) structuration theory.

At the more micro-social level of interaction, talk plays a crucial role in socialising new generations to professions and academia itself: we pass on explicit and tacit understanding of the norms of academic discourse, and of preferred ways of talking. I have earlier (Mauranen 2001, 2004) compared the role of talk in academia to Gilbert and Mulkay's (1984) 'contingent' repertoire that scientists engage in informally and behind the scenes, where mundane matters of serendipity and luck get talked about along with power, interpersonal relations, and struggles over position and financial resources. This can be pitted against the 'empiricist' repertoire of written presentation, where reports of experiments, results, and their theoretical implications are presented in an impersonal and detached manner. There is more interpersonal engagement than first meets the eye in research articles, as EAP research has been keen to show over the last twenty years, but what gets written and published is still far from the intimacy and freedom of the spoken word. Most of the uses that language is put to in academia are being carried out in ELF all over the world, and it is the spoken mode that shows the first signs of change in language. For signs of new developments in academic English, speech is what we should turn to.

Academic communities using English as their lingua franca span a broad spectrum of objectives, duration, and location, ranging from research project teams to master's programmes and short exchanges of students or staff. International research project teams of a global reach may be funded for a few years at a time, be located either in their respective institutes or in one location, or divided between different arrangements. Some research centres recruit internationally but are permanently located in one place (the Max Planck Institute; CERN). Doctoral students and postdoctoral researchers may spend up to a few years in such teams, changing places as their careers take shape; they participate in the transnational flows that increasingly characterise the current stage of global mobility. Master's programmes of one or two years have been

mushrooming over the last decade or so, and shorter student exchange programmes at undergraduate or graduate level have become routine at least in Europe. All this mobility and its associated multi-layered networks contribute to great complexity in linguistic settings.

If speaking is crucial to EAP, why should ELF research take an interest in academic language in particular? For a number of good reasons. To begin with, academic language exerts considerable normative influence on standard languages. We are used to thinking of the "educated native speaker" as the ideal speaker that language standards are modelled on, and university is of course the institution that generates such speakers. In view of the way English is developing in the world, the target speaker may not be a *native* speaker in the future, but probably *educated* all the same. From a purely linguistic point of view, the emphasis in higher education on English all over the world brings English into contact with a very large number of the world's languages, as Thomason observed some years ago (2001). Since language contact is a major factor in bringing about linguistic change, academia provides an important source of ELF features.

Despite the straightforward aim in the first EAP corpora of reaping benefits from native speakers' language to provide a model to non-native speakers, changes had taken place in the conceptualisations of English by the time the corpora were completed. There was more awareness of cultural variability and more concern with identities. There was also budding awareness of English as an international lingua franca, a viewpoint that had been strongly put forth by scholars like Widdowson (1994), Jenkins (2000, 2007) and Seidlhofer (2001). These signs of the time found their way to the MICASE corpus, where the proportion of non-native speakers is comparatively large (12%) as a consequence. Things started moving fast at the turn of the millennium, and the first corpus of academic English spoken as a lingua franca (ELFA) began its recordings in 2001, close on the heels of the first ENL speech corpora. It is interesting to note that in the case of ELF research the usual progression from written to spoken language has been reversed; another ELF corpus, VOICE, compiled in Vienna (www.univie.ac.at/voice) and others are starting in different parts of the world, but there is no written database of English as a lingua franca as yet – although the WrELFA corpus of Helsinki (http://www.helsinki.fi/englanti/elfa/wrelfa) is now breaking new ground.

3. The ELFA corpus

English as a Lingua Franca was a virtually unexplored territory at the beginning of the millennium, and academic ELF a completely white spot on the map when the compilation of a corpus of academic ELF speech was begun in Finland in 2001.

Considering ELF from the point of view of corpus compilation, it might seem at first glance that a general reference corpus would be the most desirable database for exploring a new use of English. However, in sheer practical terms it is hardly a manageable task; a project seeking to capture a representative corpus of a global language would require enormous resources. A more feasible approach is international collaboration following the models set by the *International Corpus of Learner English* (ICLE; www.uclouvain.be/en-cecl-icle.html), or the *International Corpus of English* (ICE; http://ice-corpora.net/ice/index.htm): collaboration between teams of researchers from different countries. Anything less would inevitably suffer from limitations of local features.

Another route to making the task of an exploratory corpus more manageable is to narrow it down as an alternative to expansion: focus the whole effort on a key area that can be delimited and investigated reasonably reliably. A specialised corpus is able to maintain a clearer focus on its domain and thereby of the questions that can be put to the data, yielding a clearer interpretation of findings. In effect, focus and collaborative international teamwork can be achieved at the same time, as shown by the ICLE corpus, which has collected data that is clearly delimited genre-wise. For ELF, an academic corpus is well motivated, as discussed above.

The ELFA corpus was completed in 2008, and consists of 1 million words (131h of recorded speech) of spoken English in university contexts. It is accessible to all interested researchers. The compilation principles and the choices made are discussed briefly in the following two sections, and described in more detail in Mauranen (2006b), Mauranen and Ranta (2008) and Mauranen et al. (2010).

3.1 Setting-related choices

ELFA compilation principles are essentially 'external', that is, the prominent genres of the discourse community have been identified on a social, not language-internal basis. The speech event types reflect the naming practices of the relevant discourse communities, reflecting their self-understanding of their activities. In this way, data gathering was informed by 'local knowledge' (Geertz 1983), based on informal interviews and publicized material (such as websites) of the communities about themselves. The 'folk genres' identified in this way were those that actors such as faculties, departments, or conference organisers had identified and named as their own activities. Many of the resulting event labels like "seminar" and "thesis defence" were used across the institutions. In this way, the corpus is relevant to its social setting, and has social grounding in the communities of practice where its speech events are being used and regulated.

The basic unit of sampling was the 'speech event type', following MICASE. This is a looser term than 'genre', and therefore perhaps more appropriate, as some of the event types were more firmly established as genres across the board (e.g., lectures) than others (e.g., panel discussion). The commonest event types were more central in their institutional contexts. Typicality played a role: event types that many disciplines and departments shared – seminars, lectures, thesis defences – were taken to represent the regular activities going on in the relevant discourse communities, and therefore deemed important for inclusion in the corpus. Such events are also influential in that they concern a large number of people in the institutions. Conference presentations and discussions are obviously more relevant to academic staff than to students, but nevertheless significant in academic practices. All these event types were included, with typical university-internal discourses and international programmes at the centre. The distribution of the event types is shown in Figure 1.

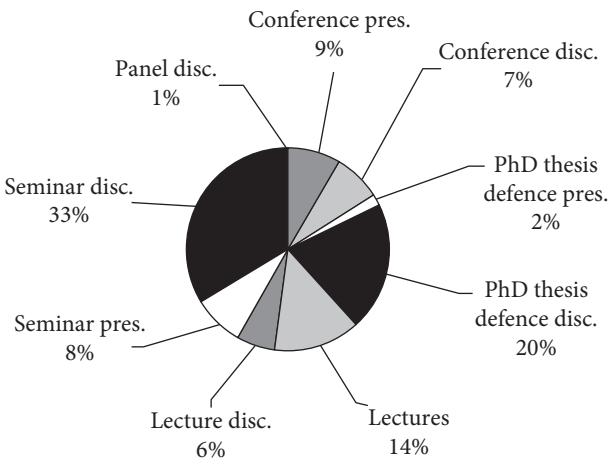

Figure 1. Distribution of event types in the ELFA corpus.
Abbreviations: pres. = presentations, disc. = discussions (from Mauranen & Ranta 2008)

Both of the pioneering U.S. corpora of academic speech focused on one university at the outset. T2K-SWAL has since branched out, and it was not exclusively a speech corpus to begin with. MICASE was compiled deliberately with one university in mind. This made good sense because a large university with wide disciplinary coverage is arguably as good an estimate of representing academic speech genres as any other, barring an extensive research-based description of the kinds and distributions of university genres on average. Corpora are not normally based on such research, which would consist in separate projects preceding actual corpus compilation.

This is an issue that make corpora an easy target for criticism concerning their representativeness: while they may have sound principles for effective and expedient compilation in language related terms, they rarely prioritise purely statistical considerations. Categories are more often based on the compilers' notions of language, genre, or register, combined with stratified sampling techniques. The resulting databases therefore reflect language experts' views on the relevance of text types, as well as the compilers' particular theoretical stances – even if these are mostly left implicit. Large reference corpora usually seek to cover as much as possible of the language of their time, so that in addition to size, wide coverage is a central target. In this way, corpora tend to reflect the prevalent notions of language at their time of compilation; they are subject to ageing from the conceptual viewpoint as theoretical frameworks change in the field. This adds a facet of ageing to the more obvious changes of language itself and the development in technological possibilities of compilation.

Compiling an ELF corpus in an academic environment where English is not universally used as a language of teaching or administration cannot assume the same overall event type selection as a single-university corpus in an ENL context. In terms of genres and disciplinary coverage, the possibilities are considerably narrower because English-medium programmes are not evenly distributed across departments and disciplines. The compilation of ELFA began at Tampere, a university whose profile is strong on social and behavioural sciences, medicine, and arts, but lacks for example a science faculty. It was therefore felt that disciplinary areas from other universities should be included so as to get a better-balanced selection of disciplinary areas into the corpus. As can be seen in Figure 2, the balance is still somewhat tipped in favour of social sciences, but the coverage is nevertheless wide, comprising both 'soft' and 'hard' sciences, and distinctly broader than a single institution would have offered.

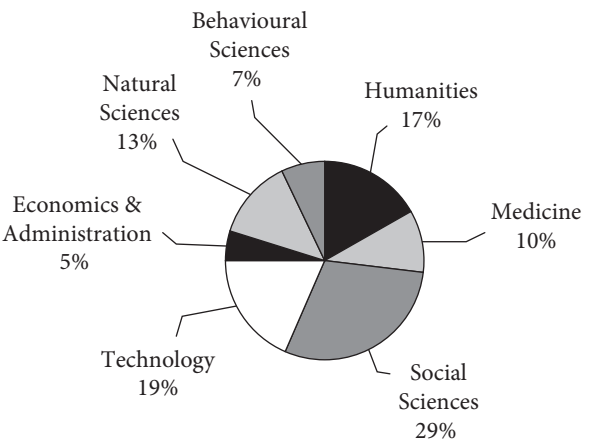

Figure 2. Distribution of disciplinary domains in the ELFA corpus (from Mauranen & Ranta 2008)

This is obviously a compromise between the reality of a given institution and some conceivable ideal balance of disciplines, if indeed such an ideal can be specified. Universities have different disciplinary profiles, and identical departmental labels do not reflect identical divisions into disciplines or subdisciplines (see e.g., Mauranen 2006c). Thus even if aggregate information of the disciplinary distributions of all the world's universities were available, it might not be a reliable guide to the kinds of academic activities actually being carried out. ELFA opted for an 'improved' reality in rounding out the corpus to include a wider selection of disciplines than one or two universities would have yielded on their own. In this way, the disciplinary selection followed the approach adopted in large reference corpora of including something of every major genre. It follows that caution needs to be exercised in inter-generic and interdisciplinary comparisons – the data represents academic discourse as a large aggregate body, while its individual components do not claim to represent that particular discipline or genre in a balanced way.

3.2 Speaker-related choices

Speakers that use English as a contact language amongst them are not learners of English, which is why an ELF corpus must be clearly distinguished from a learner corpus such as the ICLE (www.uclouvain.be/en-277586.html) (for a more detailed discussion, see Mauranen 2012). Obviously, a number of linguistic features are shared between learners and ELF speakers, and people can alternate in both roles even during the same day. Still, there are very strong reasons for keeping learner and speaker events apart, because social, cognitive, and interactive parameters shift in important ways when we move from one of these event types to another.

If a corpus targets authentic language use, situational parameters must reflect this as closely as possible. A crucial difference between learner and ELF corpora is that learner corpora keep a close eye on the proficiency level of the learners in the corpus or any section of it. This makes sense in view of the questions asked of learner corpora, which often relate to stages in L2 development. In contrast, attempts to keep proficiency constant would be counterproductive for an ELF corpus, because ELF is commonly used between speakers of varying proficiencies. Attempts to control for proficiency would miss out on an important situational parameter, the natural asymmetries among speakers. In sum, the corpus consists of naturally occurring situations where English is the real lingua franca, where participants may have different proficiencies, do not share a L1, and where they are not in an ELT class.

Despite the general aim of ELFA to prioritise external compilation criteria, some criteria are nevertheless language-internal – such as the speaker-related criterion of linguistic background. The objective was to get as much variation in the

speakers' language background as possible, and to keep the proportion of Finnish L1 speakers below 50%. Both goals were successfully met: 51 typologically highly diverse first languages are represented, and the proportion of Finns is a little over a quarter (28%).

The second question concerning speakers and their linguistic background relates to ENL speakers. The role of ENL speakers in ELF has been much discussed, and some scholars keep to the narrow definition of lingua franca stipulating that English is not the native language of *any* of the participants (also e.g., Firth 1996; Meierkord 1998). A broader definition following Thomason (2001) sees it as a vehicular language spoken by people who do not share a native language (e.g., Mauranen 2003). While the broad definition is more realistic in terms of commonly occurring speaker combinations, it has the downside of having to draw the line somewhere between a L1–L2 conversation and a lingua franca conversation. This is not easy, although an intuitively satisfying solution is that a dyadic L1–L2 conversation is the limiting case. Going on from that, the decision of VOICE to include only situations with less than 50% of ENL speakers is satisfactory. The proportion of ENL speakers in any ELFA recording comes nowhere near this limit, their total proportion being 5%. This is less than half of the NNSs in MICASE, who account for approximately 12% of the corpus. None of the ENL speakers in ELFA appear in prominent roles such as lecturers, presenters, or examiners. Their roles are relatively peripheral, largely limited to participation in discussions.

The second essentially language-internal criterion was a deliberate bias for dialogic events. Again, this was not a self-evident choice. Lectures and presentations feature prominently in universities, and their large proportion could be defended on that account. Conference and student presentations also make up important goals for academic novices to master. Many research questions about accents and typical non-standard features can also benefit from monologues just as well as from dialogues. Moreover, monologues are far easier to record and transcribe, meaning they can reach higher reliability and accumulate words into a corpus at a faster pace. Despite these advantages, monologues are not able to provide answers to certain questions crucial to understanding language change and linguistic self-regulation in groups; since it can be reasonably argued that interactional discourse is the most fundamental form of language, it provides the best context for observing language and norms in the making. Accommodation at all levels of language is one of the most intriguing aspects of linguistic interaction and likely to hold the key to understanding both conventionality and change. It is also in interaction that the different linguistic and cultural backgrounds come together and negotiate their differences and commonalities. Crucially, interaction is the only situation where miscommunication may surface: monologues may or may not be understood by their hearers, but there is no way in which this may be ascertained from the speech

data. Questionnaires are able to give only very indirect information about this. The dialogic bias thus essentially relies on the theoretical research interests concerning the corpus, motivating also these speaker-based and language-internal compilation principles.

In addition to such basically theoretical issues, practical matters impose their own restrictions. Despite the best efforts of compilers, and good principles laid down for the ideal corpus, reality tends to get in the way of achieving all the goals set. Opportunistic sampling based on the data that is actually accessible cannot completely be avoided however good the planning: not all permissions are obtained, not all recordings are successful. The limitations of compilation have consequences on the conclusions we can draw from corpora. Since a major advantage of corpus-based study over small-scale qualitative research is that they allow a far better understanding of frequencies and preferences in the language, representativeness issues are crucial to our claims.

Issues of representativeness in ELFA are solved much along the lines of general reference corpora: the database represents speaking in certain settings, and seeks to include as wide a variety of the speech event types, disciplines and language backgrounds as possible. The purpose is to secure general coverage in the specific dimensions, so as to allow searching for commonalities within a given domain, and compare it to others as becomes relevant. In contrast, an approach of this kind does not permit reliable intra-corpus comparisons between genres, disciplines, or language backgrounds. The representativeness of any of these subsets works only as part of the relevant larger whole, which is the horizon of the domain- and mode-specific corpus.

ELFA is a relatively small corpus. While corpus size has been multiplying all along with the advent of new technological tools, specialised corpora still remain comparatively small. Specialisation offsets some limitations of size, given that special corpora address more focused research questions. Corpora of spoken language also tend to be nowhere near the size of written corpora especially in the age of Internet downloading. The much-used *London Corpus of Spoken English* (Svartvik 1990) comprises only half a million words. All things considered, ELFA is a corpus of a substantial size.

In brief, then, speech event types in ELFA result from considering discourse communities' self-perceptions, a wide disciplinary selection, relevant speaker attributes and the kinds of research questions the database was set up to answer. It provides a basis for charting linguistic territory on a wide front, because it is sufficiently large and balanced in terms of first languages and speech event types to enable exploratory studies in a new research field. It is robust enough for testing and generating hypotheses on lingua franca in academia.

4. Using corpora of professional speaking

Research based on large electronic corpora has become a normal part of many kinds of linguistic enquiry, not only the branch of linguistics known as 'corpus linguistics'. The latter is nevertheless also very much alive as an approach to linguistics in its own right including a number of controversies that have sprung up around it. About three years ago, the *International Journal of Corpus linguistics* (Worlock Pope 2010) dedicated an entire issue to some of the ongoing debates, such as whether corpus linguistics is a theory, a method, an approach or something else. While this is not the place to engage in the full debate, some of the issues have particular bearing on the ways in which we may make the best use of corpus data. One relates to the distinction originally drawn by Tognini-Bonelli (2001) between 'corpus-based' and 'corpus-driven' approaches. In essence, it is analogous to the inductive vs. deductive distinction in general research methodology, with similar problems if taken to extremes. However, it is not fruitful to seek to equate 'corpus-driven' with intuitive analysis as is done by Gries (2010) any more than to equate 'corpus-based' with more rigorous theoretically motivated analyses, as seems to be suggested by Xiao (2009) and again by Gries (2010). In the following sections, I take some practical examples to illustrate the problematic nature of such simple distinctions, and in line with the topic of this paper, I draw them from corpus study of academic speech, primarily ELFA, but also occasionally making use of MICASE for comparison.

4.1 Starting by brainstorming

I set out with a case that adopts a very common approach to corpus utilisation: taking a functional category as a point of departure, selecting manifestations of that category, and searching a corpus to discover frequencies and distributions of this array of expressions.

This case is concerned with "Announcements of Self-Repair" as investigated by Marx and Swales (2005), a good case in that it is published on the MICASE website and represents a methodological approach much used in studies based on that corpus. This is how they explain their approach:

> We began by brainstorming phrases that a speaker might use when he or she wanted to tell the interlocutors that an attempt to fix a speech mistake, clarify an idea, or rephrase an ambiguous utterance was coming up. (Marx & Swales 2005)

It is clear from this self-description that the approach is not corpus-driven, and that it is intuitive. More relevantly to the present case, it is important to distinguish

between on the one hand intuition as referring to intuitive, undefined, or not very rigorously defined categories, which is a common feature in many inductive methods, and on the other hand intuition as referring to native speaker intuition, which is specific to linguistics. In the above statement, Marx and Swales provide a relatively loose definition of the category they had in mind, and it might be criticised for lack of rigour. Be that as it may, a more serious issue from the present point of view is that they do not address the second sense of intuition at all, the fact that they resorted to their linguistic intuitions as native speakers of English. The tacit assumption is that the native speaker's intuition is a reliable guide to the array of expressions available for a given function in the language.

Before continuing with the argument, let us first look at the outcome of their search (Table 1).

Table 1. Announcements of self-repairs in the MICASE corpus (Marx & Swales 2005)

Expression	N	/ 100,000 words
in other words	224	12.1
(I) mean	50	2.7
trying to say	19	1.0
another way	18	1.0
that is to say	16	.9
namely	15	.8
i.e.	14	.8
meant	11	.6
what I'm saying is	7	.4
clarify	4	.2
rephrase	4	.2
more specifically	2	.1
misspoke	1	0
Total	385	20.8

As Table 1 shows, there is a very strong preference for one pattern, namely *in other words*, followed by a minor preference, *I mean*, and after that the usage disperses among small, infrequently occurring items. The method that produced this result has certain advantages: most importantly, it provides a convenient shortcut to corpus searches. Moreover, it does not distort the figures. Each expression is found by the corpus search reliably and reflects faithfully their frequencies and their distribution. For purposes of direct application to teaching or translation for instance, this may look like a very useful, straightforward method.

However, there are also serious caveats. First, a method of this kind dismisses the possibility of finding something that a native speaker (even two) cannot think