Corpora in Applied Linguistics

Corpora in Applied Linguistics:

Current Approaches

Edited by

Nikola Dobrić, Eva-Maria Graf
and Alexander Onysko

Cambridge
Scholars
Publishing

Corpora in Applied Linguistics: Current Approaches

Edited by Nikola Dobrić, Eva-Maria Graf and Alexander Onysko

This book first published 2016

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-9464-8
ISBN (13): 978-1-4438-9464-7

CONTENTS

Introduction 1
Nikola Dobrić, Eva-Maria Graf and Alexander Onysko

Research on L2 Pragmatics at a conceptual and methodological interface 9
Marcus Callies

A focus on pragmatic competence: The use of pragmatic markers
in a corpus of Business English textbooks..... 33
Peter Furkó

Written summarisation for academic writing skills development:
A corpus based contrastive investigation of EFL student writing 53
Gyula Tankó

Refining the Scope – Substance error taxonomy: A closer look
at Substance..... 79
Günther Sigott, Hermann Cesnik and Nikola Dobrić

The uses and functions of metadiscourse in intercultural project
discussions on language education 95
Hermine Penz

Does code-switching exist in personal writing? 121
Olga Grebeshkova

Frequency analysis of trigger words and money-related expressions
in British and Serbian bank offers 145
Vesna Lazović

"I use English, but if need be I'm fluent in German as well":
Croatian business professionals' use of English and other languages 165
Branka Drljača Margić and Irena Vodopija-Krstanović

INTRODUCTION

NIKOLA DOBRIĆ, EVA-MARIA GRAF
AND ALEXANDER ONYSKO

Corpora in Applied Linguistics – Current Approaches brings together contributions from the Klagenfurt Conference of Corpus-Based Applied Linguistics (CALK14). The volume aims to extend corpus linguistic research in different areas of applied linguistics. As such, the articles in the book discuss diverse areas of applied linguistics research based on authentic language data and corpora.

Applied Linguistics – broadly defined as a scientific approach and means to help solve language-related problems in society – is a branch of linguistics that benefits substantially from using corpora in its research. Even though the use of corpora has recently led to the emergence of a separate field of Learner Corpus Research (e.g. Callies & Götz 2015; Granger, Guilquin & Meunier 2015), other areas of applied linguistics have not yet explored the practical potential of corpora to their fullest extent. Besides the by now classic work of Susan Hunston (2002) on *Corpora in Applied Linguistics*, there are only two more edited volumes that specifically deal with the use of corpora in applied linguistics, i.e. Campoy and Luzón's (2007) volume on *Spoken Corpora in Applied Linguistics* and Hyland, Meng Huat and Handford's (2012) volume on *Corpus Applications in Applied Linguistics.* Thus, the current volume would like to draw some more attention on the vast possibilities of applying corpora in carrying out research in applied linguistics.

The use of corpora is most closely linked with the field and method of corpus linguistics and its commitment to using empirical 'real' language data for linguistic research. The field and its methods have evolved and developed immensely over the last 50 years. According to Hunston (2002: 1), "[i]t is no exaggeration to say that corpora, and the study of corpora, have revolutionized the study of language, and of the application of language, over the last few decades."

Contemporary corpus linguistics, which is based on the computer-aided analysis of large databases of text, had its onset in the late 1950s and

more concretely in 1961 when the texts for the Brown Corpus were collected (Francis & Kucera 1964). This first modern, computer readable, general corpus of Standard American English contains one million words of American English texts from 15 different text categories. While the Brown Corpus set the standards for corpus compilation for years, the release of the British National Corpus (BNC) in 1995 represented another milestone by its sheer size of 100 million words. Nowadays, electronic accessability of texts particularly on the world wide web facilitate the compilation of large-scale corpora, such as the NOW corpus compiled by Mark Davies, which includes English web-based newspaper contents since 2010 and grows in size by 4 million words every day. Apart from the English language (and its varieties), corpora have also become important for the study of many languages, including languages with smaller numbers of speakers, for which corpora can be important tools of language documentation, maintainance and revitalization (e.g. Boyce 2006). In addition, the last two decades of corpus linguistic research have also experienced a trend towards smaller, custom-made corpora that are targeted for answering specific questions. The contributions in the current volume are a further testimony of that.

Taking another brief look back at the development of corpus linguistics, it becomes clear that corpus-based and corpus-driven approaches to language analysis originated from the struggle between studying idealized speaker utterances, and thus relying on intuitions, and considering actual language use and variation in an empirical fashion. Since the 1980s, corpus linguistics has had a constantly increasing impact on the study of language due to the following two factors: The first relates to the growing awareness that intuition alone is not a sufficient and adequate method for providing solid linguistic evidence. The second factor is spawned by the information revolution and the development of ever more powerful and affordable computers as well as the rise of the Internet and new information technologies.

Today, there is a plethora of corpora of various types geared towards answering different kinds of linguistic and didactic questions. Methodologies of corpus-based research have been thoroughly developed, including computer software and statistical methods for dealing with quantitative data. The role of corpora is by now universally recognized as essential in contemporary linguistics. This is not only true in certain areas of linguistic exploration such as lexicography or natural language processing, but also other branches of linguistics are increasingly applying corpus-based methods in their empirical practice. We hope that this

volume helps to further disseminate the benefits and advantages of working with corpora in the vast field of applied linguistics.

The studies gathered in this volume explore the opportunities that both spoken and written corpora offer for answering questions in different domains of applied linguistics such as second language learning, language testing, comparative linguistics, learner pragmatics and specialized discourses. At the same time, the contributions also give insight into possible limitations and further challenges of corpus-based research in these areas.

In more detail, the opening contribution by Marcus Callies on *Research on L2 Pragmatics at a conceptual and methodological interface* addresses the construct of pragmatic knowledge in a foreign/second language (L2) at the conceptual interface of pragmatics, syntax and discourse, and the methodological interface of Second Language Acquisition (SLA) research and Learner Corpus Research (LCR). The study focuses on the pragmalinguistic component of L2 pragmatic knowledge by examining a means of information highlighting, i.e. a specific type of cleft construction. It argues that both SLA research and LCR can benefit from the possibilities offered by the use of learner corpora in the study of interface relations in language acquisition: corpus and experimental data could be used as potentially converging evidence when studying interface phenomena. A case study of learners' use of demonstrative clefts exemplifies the crucial role of spoken learner corpora in the expansion of the research agenda of interlanguage pragmatics in that they enable researchers to study a much broader range of different pragmatic phenomena on the basis of authentic, continuous and contextualized data.

Peter Furkó's article, *A focus on pragmatic competence: The use of pragmatic markers in a corpus of Business English textbooks*, looks into the role of pragmatic markers (PM) in shaping ESL speakers' communicative competence in the context of business communication. Despite their essential role in organizing and structuring discourse and for marking speakers' attitudes towards the propositional content of their utterances, elements such as *well, you know, of course, right*, etc. take a back seat in TEFL, TESL and, most notably, in TESP contexts. The paper discusses the major issues related to the concept of communicative competence. In particular, the role of PMs in ESL is addressed, including difficulties that may hinder learners from acquiring the proper use of PMs. A case study on the representation of PMs in Business English textbooks aims at mapping the functional spectrum of PMs in the selected teaching materials. The findings show that attention towards the importance of PMs

has increased in Business English textbooks over the last ten to fifteen years, which is evident by the presence of explicit instructions and exercises on using PMs. However, if Business English textbooks are used as a corpus of utterances, the inadequate treatment of PMs becomes obvious in view of their frequency and functional range. The paper concludes by stating that teachers need to compensate for the inadequate input provided in textbooks of Business English.

Gyula Tankó's article, *Written summarisation for academic writing skills development: A corpus based contrastive investigation of EFL student writing*, presents findings from a corpus-based study that compares the effect of academic essay and summary tasks on the written production of EFL learners. The study is contextualized in both the global and local concerns for testing academic English proficiency. As such, it contributes to the findings of first studies that investigated the effect of task type on written language production. The author analyses the syntactic and lexical characteristics of two writing tasks, an independent argumentative essay and an integrated guided summary writing task with the aim to determine whether the task types elicit written academic English discourse. Furthermore, the study considers the practical implications of the findings for EAP teaching and assessment. The results show that the argumentative essay and guided summary tasks elicit language with the characteristic features of written academic prose. At the same time, there are marked differences in the participants' written production on the two tasks. Based on this evidence, Tankó formulates a few recommendations for EAP teachers and assessors with the proviso that the educational application of these results may require fine-tuning according to the varying rhetorical traditions of academic disciplines and the impact of the students' L1 background.

The next contribution by Günther Sigott, Hermann Cesnik and Nikola Dobrić, *Refining the scope – Substance error taxonomy: A closer look at Substance*, deals with the highly complex task of identifying and describing errors in learner language, in particular in L2 writing. The paper, which is an elaboration of the *Scope – Substance* error taxonomy developed in Dobrić & Sigott (2014), aims to develop a methodology for recording annotator agreement and to determine the degree to which annotators agree on the *substance* of errors after being introduced to the principles underlying the *Scope – Substance* error taxonomy. *Scope* refers to the amount of context that is necessary in order for an error to become perceptible. *Substance*, by contrast, refers to the amount of text that needs to be changed so that the error will disappear. Since agreement on error *Substance* is a prerequisite for agreement on error type, which results from

the combination of *Substance* with *Scope*, the present study focuses only on *Substance*. The degree of agreement reached and the problems encountered are discussed as the basis for further refining the *Scope – Substance* error taxonomy towards future application in corpus annotation, teaching and assessment.

The contribution on *The uses and functions of metadiscourse in intercultural project discussions on language education* by Hermine Penz investigates spoken metadiscourse during intercultural project discussions. The study contributes to fostering research on spoken metadiscourse, which, unlike its written form, has received very little attention so far. The paper argues that the main function of metadiscourse in the intercultural data at hand lies in achieving understanding and creating a common basis in a context of great diversity. However, when comparing the use of metadiscourse in group discussions, which essentially could be considered as the same activity type, differences in frequency, in particular in connection with specific functions, are identified in the corpus. Such differences in frequency and types of metadiscourse are interpreted to reflect the interaction within events which could be classified as the same speech activity. On the premise that speech activities are not clear cut events and are characterised by fuzziness, the author concludes that an analysis of metadiscourse can help to uncover variation within activity types (or genres). This is indicative of the fact that metadiscourse can be seen as a reflection of the socio-pragmatic context.

Olga Grebeshkova's chapter, *Does code-switching exist in personal writing?*, summarizes the first findings from an on-going study of second-language learners' personal writing, and in particular of the phenomenon of code-switching. The author sets out to answer the question to what extent personal writing, i.e. writing from an author that is addressed to the same author, in the bilingual environment is affected by language switches. In order to do so, she has built a corpus that consists of 83 examination notes from French students taken during their Bachelor degree examinations in English and of 83 examination notes from Russian students taken during their final English exam in the 4th year of their studies. Out of these notes, 43 contain instances of code-switching, which the author further analyses on the basis of language content relationships of multilingual texts and the use of intra-/inter-sentential code-switching. As a preliminary result, the data in Grebeshkova's research manifest the presence of code-switching in personal writing as concrete, genuine evidence of bilingual writing.

Vesna Lazović's contribution, *Frequency analysis of trigger words and money-related expressions in British and Serbian bank offers*,

explores the use of trigger words and money-related expressions in British and Serbian bank offers. Based on data collected from web pages of banks operating in two countries and offering services to people of two different cultures, the study highlights current trends in advertising in different languages at the lexical level. In particular, it emphasizes the ways words are used to persuade, convince and manipulate potential clients. Furthermore, it compares the frequency analysis of those expressions in English with their translation equivalents in Serbian in order to, first, demonstrate whether different cultural background influences advertising messages and, second, to reveal similarities and differences in the lexical approach to financial products. The data analysis confirms that some features in advertising are universal in different languages and cultures, as is the use of trigger words and money-saving expressions. Yet, the study also demonstrates that culture can have a significant influence on marketing decisions and the success of marketing communications. The author concludes that more cross-cultural analyses are required to raise awareness towards cultural differences in the perception and reception of marketing messages in the Internet era. Only in that way can universals of register variation be established and linguistic methods and strategies across cultures understood.

The final contribution by Branka Drljača Margić and Irena Vodopija-Krstanović entitled, *"I use English, but if need be I'm fluent in German as well": Croatian business professionals' use of English and other languages*, analyzes the use, status and importance of English and other languages in the Croatian business environment. The larger socio-cultural context of the study is Croatia's recent accession to the European Union and the fact that its membership in the single market of the European Union has offered new opportunities for business networking and has affected the use of languages for business purposes. The study focuses on Croatian business professionals' self-reported use of languages and their stance towards the importance of English and other languages. The findings show a consensus about the strong convergence towards English in the Croatian context; English is considered to be the most significant language in the business domain although other languages are also deemed useful. In general, the study sheds light on the way in which English is used as a business lingua franca, and on how English and other languages can contribute to ensure greater success in the global marketplace. Although the study is not concerned with teaching, the authors conclude that their findings could have implications for English language training programmes for business purposes, teaching of English for specific

purposes at university, and for English-medium instruction in business programmes in higher education.

References

Boyce, M.T. (2006). *A Corpus of Modern Spoken Māori*. Ph.D. thesis. Victoria University of Wellington.

British National Corpus (BNC). University of Oxford. http://www.natcorp.ox.ac.uk/; last accessed May 16, 2016

Callies, M. & S. Götz. (eds.) (2015*). Learner Corpora in Language Testing and Assessment*. Amsterdam & Philadelphia: John Benjamins.

Campoy, M. C. & M. J. Luzón (eds.) (2007). *Spoken Corpora in Applied Linguistics*. Bern: Peter Lang.

Davies, M. *NOW Corpus (News on the Web)*. Brigham Young University. http://corpus.byu.edu/now/; last accessed May 16, 2016.

Dobrić, N. & G. Sigott (2014). Towards an error taxonomy for student writing. *Zeitschrift für interkulturellen Fremdsprachenunterricht* 19 (2): 111–118.

Francis, W.N. & H. Kucera (1964). *Brown Corpus Manual*. Brown University. http://www.hit.uib.no/icame/brown/bcm.html; last accessed May 16, 2016.

Granger, S., G. Guilquin & F. Meunier (eds.) (2015). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Hyland, K., C. Meng Huat & M. Handford (eds.) (2012). *Corpus Applications in Applied Linguistics*. London: Bloomsbury Publishing.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

RESEARCH ON L2 PRAGMATICS AT A CONCEPTUAL AND METHODOLOGICAL INTERFACE

MARCUS CALLIES

1. Introduction

This chapter addresses the construct of pragmatic knowledge in a foreign/second language (L2) at the conceptual interface of pragmatics, syntax and discourse, and 2) at the methodological interface of Second Language Acquisition (SLA) research and Learner Corpus Research (LCR). Its aim is to contribute to the growing number of studies that use spoken learner corpora to study features of the grammar of conversation. The study focuses on the pragmalinguistic component of L2 pragmatic knowledge by examining a means of information highlighting, i.e. a specific type of cleft construction.

2. The conceptual interface. Pragmatics in Second Language Acquisition: Going beyond speech acts

Callies (2013) argued that the study of pragmatics as a field of inquiry within Second Language Acquisition (SLA), usually referred to as Interlanguage Pragmatics (ILP), has traditionally adopted a narrow research focus, and that the significance of L2 pragmatic knowledge beyond the domain of speech acts has been neglected in ILP to date. Similarly, Bardovi-Harlig's (2010) state-of-the-art meta-analysis of published research in ILP concluded that "the dominant area of investigation within interlanguage pragmatics has been the speech act" (2010: 219). However, recent developments suggest that there is a growing awareness in the field that L2 pragmatics is more than speech acts and that the scope of inquiry needs to be adjusted accordingly (LoCastro 2011: 333).

Pragmatic knowledge in an L2 clearly includes more than the sociopragmatic and pragmalinguistic abilities for understanding and performing speech acts. Standard descriptions of ILP frequently use notions like "linguistic action in L2" (Kasper 2010: 141) to refer to the general domain of inquiry. Definitions of pragmatic knowledge or competence[1] range from rather broad and general ones, e.g. "the ability to use language appropriately in a social context" (Taguchi 2009: 1) to more detailed ones, e.g. "the knowledge of the linguistic resources available in a given language for realizing particular illocutions, knowledge of the sequential aspects of speech acts and finally, knowledge of the appropriate contextual use of the particular languages' linguistic resources" (Barron 2003: 10).[2] Callies proposed the following definition of pragmatic knowledge:

> L2 pragmatic knowledge is the knowledge of the (pragma-) linguistic resources available in a particular language for realizing communicative intentions, and the knowledge of the appropriate socio-contextual use of these resources. Pragmalinguistic knowledge is a component of L2 pragmatic knowledge which relates to learners' knowledge of the structural linguistic resources available in a given language for realizing particular communicative effects, and knowledge of the appropriate contextual use of these resources. (2013: 14)

There are a number of models of language proficiency that aim to capture the ability of L2 learners to use language in social interaction, all of which acknowledge to some degree the importance to acquire pragmatic competence in L2 learning (see Callies 2013 for discussion). The present chapter adopts a componential view of linguistic knowledge, use and development as well as L2 proficiency that has similarities to a modular view which presupposes that a learner's knowledge of a foreign language consists of modules such as phonology, morphology, syntax, semantics and pragmatics.[3] These modules have their individual structural and functional properties. They interact with each other and with other cognitive systems. These interactional processes are known as interface relations. Interface relations have received a great deal of attention in

[1] The two terms are frequently used interchangeably in the literature.

[2] While Barron's proposal draws a useful distinction between pragmalinguistic and sociopragmatic knowledge, it centers around the concept of illocutionary acts, thus narrowing down the scope of pragmatic knowledge to sociopragmatics.

[3] Note that the view adopted here does not imply a commitment to a generative approach to SLA, although modularity is of central importance in this framework.

contemporary generative-linguistic approaches to SLA theory.[4] In these approaches, interfaces involve interactions or mappings between linguistic modules or representations. These include external ones, i.e. those where the grammar interfaces with other domains of cognition, e.g. the conceptual-intentional and the articulatory-perceptual system (e.g. the relationship between syntax and discourse), and internal ones, i.e. where different modules of the grammar interface with each other (the interface between syntax and semantics or syntax and phonology). Note that in this view, discourse-pragmatics is considered an external interface.

A highly influential theoretical construct in the study of advanced learner language has been the so-called Interface Hypothesis (IH), most recently stated in Sorace (2011). It was originally applied to the concept of ultimate attainment at the level of near-native L2 proficiency and proposed that language structures involving an interface between syntax and other cognitive domains are less likely to be acquired completely than structures that do not involve this interface. Thus, it claims that interfaces are especially vulnerable for adult learners (more vulnerable than purely syntactic features) and therefore subject to greater difficulty and delay in acquisition. Interface relations, opaque form-meaning mappings, optionality and discourse-motivated preferences are generally assumed to be among the main areas of difficulty in advanced SLA (DeKeyser 2005). Linguistic phenomena located at the external interfaces are expected to result in greater difficulties than internal ones because properties at external interfaces draw on information across linguistic and non-linguistic cognitive modules and require more processing resources. Difficulties with interface phenomena are believed to be caused by limitations in working memory as well as processing capacity and efficiency as inherent features of bilingualism.

Outside of the generative framework, interface relations in terms of the interrelationship of grammatical and pragmalinguistic abilities in SLA has also been discussed in ILP. In the majority of studies that have been conducted, pragmatic competence is singled out as an individual component of communicative competence and, thus, treated and investigated as an independent component of grammar (Kasper & Rose 2002: 159, 163). Some authors have identified a lack of research which explores the relationship between grammatical and pragmatic abilities in SLA (Bardovi-Harlig 1999, Kasper 2001, Kasper & Rose 2002), which is still

[4] See White (2009, 2011) for overview discussions and the special issue of *Lingua* (121:4) published in 2011, entitled "Acquisition at the Linguistic Interfaces" that contains studies adopting an interface-conditioned view of mental linguistic architecture framed within generative-linguistic theory.

an unresolved issue. It is suggested that the development of pragmatic competence has to be seen as independent of the development of grammatical competence since "high levels of grammatical competence do not guarantee concomitant high levels of pragmatic competence" (Bardovi-Harlig 1999: 686).

Kasper (2001: 506) and Kasper and Rose (2002, chapter 5) summarize the research findings on the relationship of interlanguage pragmatic and grammatical development which has led to two scenarios:

- pragmatics precedes grammar: learners use L2 pragmatic functions before they acquire the L2 grammatical forms that are acceptable realizations of those functions;
- grammar precedes pragmatics: learners acquire L2 grammatical forms before they acquire their pragmalinguistic functions.

In support of the first scenario, Kasper and Rose draw on the "universal pragmatics principle" and functional approaches to SLA. A persistent belief in traditional foreign language teaching is the primacy-of-grammar view which claims that in order to successfully communicate in an L2 in terms of (socio)pragmatics, learners first need to have a solid knowledge of the target language grammar. However, the universal pragmatics principle states that unlike children in L1 acquisition, L2 learners are usually pragmatically competent in their L1, hence they bring a supposedly universal pragmatic knowledge to the task of L2 learning (Kasper & Rose 2002: 164). Moreover, functionally oriented research into the early stages of untutored SLA has found that learners move from a pragmatic mode through a process of syntacticization or acquisitional grammaticalization to a syntactic mode.

The grammar-precedes-pragmatics scenario comes in three forms (see Kasper & Rose 2002: 174ff.):

- grammatical knowledge does not enable pragmalinguistic use (for example learners' (non-)use of modal verbs in mitigating disagreement);
- grammatical knowledge enables non-target-like pragmalinguistic use (for example the overuse and pragmatic overextension of *I think*), and
- grammatical and pragmalinguistic knowledge enable non-target-like sociopragmatic use (for example learners' use of information questions as indirect strategies in a number of speech act types and contexts in which more transparent strategies would be more effective).

In sum, research findings suggest that there are differences as to the pragmalinguistic development of learners at different developmental stages in the L2 learning process. However, it still remains unclear how grammatical and pragmatic knowledge in an L2 is exactly related to each other.

3. The methodological interface. Learner corpora in SLA research

Similar to SLA research in general, in which highly controlled (quasi-) experimental data have traditionally been favoured, research in ILP has largely relied on elicited assessment and production data. The most typically used data collection technique used is the discourse completion task (DCT) to elicit pseudo-oral production data about sociopragmatic behaviour in a specific communicative context. According to LoCastro (2011: 331), this may be another reason for the dominance of research on speech acts in ILP.

Corpora and corpus linguistic tools and methods are also increasingly used for the study of SLA, in particular in learner corpus research (LCR), an interdisciplinary field at the crossroads of corpus linguistics, SLA research and foreign language teaching (see Granger, Meunier & Gilquin 2015). However, its links with SLA research in general and the impact of LCR on SLA theory in particular has been limited to date, which seems especially true for SLA research couched in a generative-linguistic framework in which data-driven, usage-based approaches are not as appealing (Granger 2009: 14). It seems, however, that the two fields are now coming closer together, with learner corpus researchers recognizing the importance of SLA theory and SLA researchers gaining insights into the added value of learner corpora (Granger 2009: 14; see also Tracy-Ventura & Myles 2015 and Lozano & Mendikoetxea to appear). For example, Rankin (2009) argues that the study of interface relations is a field that should be beneficial for both LCR and (generativist) SLA. LCR has made a substantial contribution to the description of advanced written interlanguages with a focus on lexico-grammar with some corpus-based research showing that advanced learners typically struggle with the acquisition of optional and highly L2-specific linguistic phenomena, often located at the interfaces, e.g. in discourse and information structure (e.g. Callies 2009a and the studies reviewed therein). This partially mirrors recent developments in generative SLA, where the interfaces between syntax, discourse-pragmatics and semantics have been at the center of attention.

As of yet, there are only few studies that have addressed the methodological interface between generative approaches to SLA and learner corpus research (e.g. Lozano & Mendikoetxea 2008, 2010, to appear; Rankin 2009). These have focused on discourse-conditioned word order alternations such as subject-verb inversion and preposing in the written production of advanced L2 learners of English from various L1 backgrounds. Lozano and Mendikoetxea (to appear) propose to use converging evidence in the form of corpus and experimental data when studying interface phenomena: if learners show certain kinds of knowledge or deficits at the interfaces, this should be observed in both experimental and contextualized production data. Corpus and experimental data should therefore be combined and contrasted to better account for the observed deficits at the syntax-discourse interface and determine why some interfaced properties are more problematic than others.

In ILP, learner corpora – due to their very nature of being large systematic collections of authentic, continuous and contextualized language use (spoken or written) by L2 learners stored in electronic format – can help overcome several problems and limitations posed by the dominance of data elicitation techniques to date. Not only do learner corpora enable researchers to study a much broader range of different phenomena, but they can also provide results that may be viewed as more reliable, valid, and generalizable across populations without the lack of authenticity and replicability that often arises from the use of other types of data. They can be the basis for quantitatively oriented studies that are subjected to statistical analyses and create an opportunity for between-methods triangulation and alternative views to qualitative, ethnographic studies that have been common in pragmatics in general.

In particular, the availability of spoken learner corpora such as the *Louvain International Database of Spoken English Interlanguage* (LINDSEI; Gilquin et al. 2010) has enabled researchers to study a wider range of pragmatic features of learner language in the spoken mode.[5] The LINDSEI consists of spoken data, i.e. transcripts of interviews between learners of English as a foreign language (EFL) and English native-speaker or non-native-speaker interviewers. The learners are university undergraduates in their twenties whose proficiency level ranges from higher intermediate to advanced (being assessed on external criteria such as institutional status). The LINDSEI includes subcorpora of learners from eleven mother tongue backgrounds (e.g. German, French, Italian,

[5] See the list of publications based on the LINDSEI provided by the Centre for English Corpus Linguistics in Louvain-al-Neuve, Belgium, at http://www.uclouvain.be/en-cecl-lindsei-biblio.html

Japanese, Polish, and Spanish) with 50 interview transcripts per subcorpus, i.e. a total of about 100,000 words per component. Each interview lasts approximately fifteen minutes and involves three tasks: 1) a warm-up sequence in which interviewer and interviewee talk about a set topic, 2) a free discussion, and 3) a picture description.

Using data from corpora of spoken interlanguage, it is now possible to systematically examine lexico-grammatical patterns and syntactic structures that are part of the grammar of conversation on a broad empirical basis (see e.g. Mukherjee 2009 for a study along these lines). Other studies have investigated individual pragmalinguistic units, e.g. discourse markers (e.g. Müller 2004, 2005; Aijmer 2004, 2009, 2011; Buysse 2012, 2015), modal particles (e.g. Belz & Vyatkina 2005) and tag questions (Ramirez & Romero-Trillo 2005), as well as other features of turn- and discourse structure, e.g. performance phenomena like hesitations, repetitions and disfluencies (Gilquin 2008) or filled and unfilled pauses (see e.g. Brand & Götz 2011; Götz 2013). The present paper makes a contribution to research on the grammar of conversation in learner English and focuses on information highlighting in discourse.

4. Case study

An area where pragmalinguistic devices abound and are of crucial importance is discourse pragmatics, the "general domain of inquiry into the relationship between grammar and discourse" (Lambrecht 1994: 2). More specifically, I will be concerned with a syntactic means of information highlighting located at the interface of syntax and discourse-pragmatics. This interface is often referred to as information structure or information packaging, viz. the structuring of sentences by syntactic, prosodic, or morphological means that arises from the need to meet certain communicative demands, e.g. emphasizing a certain point, correcting a misunderstanding, or repairing a communicative breakdown.[6] Information highlighting is clearly pragmatically motivated because, more generally speaking, it serves to express certain pragmatic functions in discourse, e.g. intensification or contrast. Compared to their frequency of occurrence and difficulty of acquisition, there are still relatively few corpus-based studies that have examined the linguistic means of information highlighting in

[6] Deppermann (2011) provides a recent overview of the role and relevance of pragmatics for grammar, in particular as to the structuring and packaging of information and the framing of discursive action by means of grammatical constructions such as clefts.

English interlanguage from a pragmalinguistic perspective (see e.g. Boström Aronsson 2003; Herriman & Boström Aronsson 2009; Callies 2008a, 2008b, 2009a, 2009b). More generatively-oriented studies are Lozano and Mendikoetxea (2008, 2010) who investigate how syntactic knowledge interfaces with other cognitive systems by analyzing postverbal subjects in Italian and Spanish EFL learners' written production compared to English native speakers. Their findings show that while these learners produce verb-subject constructions under the same interface conditions as native speakers, they in fact overproduce them and make persistent errors in their syntactic encoding. These findings are interpreted as supporting proposals that these difficulties stem from problems at coordinating syntactic knowledge with knowledge from other external systems, but they suggest that the nature of such difficulties is not external to the syntax. Rankin (2009) also examined the interface between syntax and discourse-pragmatics by studying verb second (V2) structures in the written production of advanced German and Dutch EFL learners. The evidence shows that the residual V2 produced by the learner groups studied is the result of a deficit at the interface rather than the transfer of L1 V2 syntax, suggesting that the nature of V2 in the learners' L1 combined with evidence from the L2 input make it difficult for them to lose the V2 constraint, which remains a persistent option after the preposing of certain constituents.

L2 learners' knowledge (that includes awareness, comprehension, and production) of discourse organization and the (contextual) use of linguistic means of information highlighting is thus still a relatively underexplored area in SLA research, as is the interplay of pragmalinguistic knowledge and discourse organization in general. Recent findings suggest that information structure management is problematic even for advanced L2 learners and that such learners have only a limited awareness of the appropriate use of lexical and syntactic focusing devices in formal and informal registers (Callies 2009a).

In what follows, I present a learner-corpus study that investigates L2 learners' use of a specific means of information highlighting in English, i.e. a specific type of cleft construction: demonstrative clefts. Three research questions will be examined:

1. Are there differences in the frequencies of use of this cleft type in the speech of native speakers of English and learners of English as a foreign language?
2. Are there differences in how native speakers and learners use this device in terms of discourse functions?

3. Are there differences between learners from different L1 backgrounds, and if so, how can these be explained?

The case study is a contrastive interlanguage analysis (CIA) based on corpora of spoken interlanguage. In a CIA, two types of comparisons are combined (see e.g. Callies 2015). First, the interlanguage of a certain learner group, e.g. German learners of English, is compared with the language of English native speakers in order to pinpoint possible differences between the two groups. This comparison is then subsequently combined with a corresponding analysis of the interlanguage produced by a second group of learners, e.g. French learners of English. For the present case studies, the learner data are drawn from the German and French components of the LINDSEI (Gilquin et al. 2010). For comparable native speaker data the *Louvain Corpus of Native English Conversations* (LOCNEC) was used. The LOCNEC contains transcribed interviews with native speakers of British English (university students at Lancaster University in the UK) aged between eighteen and thirty years. The interviews involved the same tasks, topics and stimuli that were used for the interviews in the LINDSEI. Table 1 provides an overview of the corpora.

Name	**Writers' L1**	**Professional status**	**No. of interviews**	**No. of turns[7] (only interviewees)**
LINDSEI-F	French	university students	50	5504
LINDSEI-G	German	university students	50	6051
LOCNEC	British English	university students	50	8436

Table 1. Learner corpora used in the case study

The target structures were extracted semi-automatically[8] using *WordSmith Tools 5* (Scott 2008), followed by manual inspection and filtering of false

[7] In view of the manifold problems to operationalize the concept of sentence in transcribed spoken language and thus, to count the amount of sentences in the corpora, I chose to apply the number of speech turns as a basis of comparison.

[8] To retrieve all instances of the clefts, the search involved all instances of *that* and *this* followed by a form of *be* (*'s, is, was*) and a *wh*-word (*what, when, why, where, how*).

positives. The analysis of the data consisted in a quantitative analysis of frequencies of occurrence and a qualitative study of discourse functions.

Cleft sentences are information packaging constructions that involve the splitting of a sentence into two clauses. They are pragmatically motivated and differ from their basic counterparts in that they serve to highlight a certain phrase or clause, the cleft constituent. The most common types are *it*-clefts and *wh*-clefts (also known as pseudo-clefts). There are also other types of cleft constructions such as the reverse *wh*-cleft, in which the order of *wh*- and cleft-clause is inverted. The vast majority of reverse *wh*-clefts feature the non-contrastive, non-focal deictic demonstratives *that* or *this* as the cleft constituent, see examples (1) and (2),[9] and therefore this type is also referred to as demonstrative cleft in the literature (Biber et al. 1999: 961; Calude 2008, 2009).

(1) <A> so you you did English and ling= and linguistics to: <\A>
<B> I did English and linguistics just because **that was what I was interested in** the the interest in going into film industry has only developed since I've been at university <\B> (LOCNEC)

(2) <A> so you had to cope with those kids <\A>
<B> I had to cope with those kids completely on my own with no back-up she said you know she w= she thought it was great having someone to help she said right you're gonna take half the kids ..the worst half and you're going to teach them the same lesson as I'm teaching them here's the book **this is what I want you to teach them** go off and do it for a year <\B> (LOCNEC)

When compared to other types of cleft constructions, demonstrative clefts only rarely occur in written language but are clearly the most frequent variant in the spoken mode (Collins 1991: 178ff.; Oberlander & Delin 1996: 186; Weinert & Miller 1996: 176), occurring especially often in spontaneous spoken language, i.e. conversation (Biber et al. 1999: 961; Calude 2008: 86). Of the two demonstratives, *that* is much more frequent than *this* (Oberlander & Delin 1996: 189; Weinert & Miller 1996: 188; Biber et al. 1999: 962; Calude 2008: 79). Therefore, the majority of demonstrative clefts convey anaphoric deixis as in example (3),[10] but they can also express cataphoric deixis as in (4), function anaphorically and cataphorically simultaneously as in (5), or carry exophoric deixis, i.e. non-textual, extra-linguistic reference either in the form of shared world

[9] Demonstrative clefts are given in bold print.
[10] The discourse segment(s) that the demonstrative *that* refers to are underlined.

knowledge or physical/visual presence at the time of utterance, see example (6) (Calude 2008: 87ff.).

(3) <A> so what are you doing now as a major is it linguistics or is it <\A>
<B> <X> .. I I thought I'd been accepted for Chinese and linguistics combined <\B>
<A> [mm <\A>
<B> [and **that's what they told me when I first . came here** but now they seem to think it's only linguistics <\B> (LOCNEC)

(4) <B> that we're living I mean I had my had my own flat and it's very difficult to: go from having your own flat and[<X> privacy to <\B>
<A> [and share a kitchen <\A>
<B> living in somewhere much smaller <\B>
<A> mhm <\A>
<B> but erm <\B>
<A> but I mean Graduate College is quite okay <\A>
<B> yeah I know **that's why I decided to pay a bit more** cos I thought sharing a kitchen and a bathroom with ten people <\B>
<A> yeah <\A>
<B> [I just couldn't <\B>
<A> [especially the bathroom <\A>
<B> yeah no I I really couldn't have faced that <\B> (LOCNEC)

(5) <A> and you don't live there and you you've never seen something like that before .. but you you live in Sheffield <\A>
<B> yeah <\B>
<A> it's quite a big city isn't it <\A>
<B> it is quite big yeah **that's why I came here** cos I wanted to come to somewhere smaller <\B> (LOCNEC)

(6) <B> and she doesn't . it's not really a glamorous picture <\B>
<A> mhm <\A>
<B> or anything like that .. erm the third one it looks like he's painted it again .. erm .. new hairstyle .. smiling sat up .. it makes her look more beautiful than she is <\B>
<A> mhm <\A>
<B> <laughs> and in the fourth one she's telling all her friends of that's me **that's how I look** .. things like that <\B> (LOCNEC)

In view of their relatively fixed structure, Calude (2009) argues that demonstrative clefts show characteristics of formulaic expressions, allowing only a narrow range of elements to occur in its structural "slots". Prototypically, the demonstrative *that* occurs as the initial element. The

copula *be* only occurs in simple present and simple past tense and is most commonly used in its contracted form *'s*. The copula is then most frequently followed by *what*, less frequently by *why*, *where, when* and *how* as *wh*-words in the cleft clause (Collins 1991: 28; Oberlander & Delin 1996: 187; Weinert & Miller 1996: 188). Moreover, demonstrative clefts have a distinct function in discourse as organizational and discourse-managing markers, and are typical of a specific register, i.e. conversation.[11]

Demonstrative clefts have multiple functions as to discourse organization and management. In particular, what sets them apart from other cleft types is their pointing function by means of the initial demonstrative pronoun (Weinert & Miller 1996: 188; Oberlander & Delin 1996: 189). They typically have extended text reference that spans over three or more turns prior to the cleft (Calude 2008: 79f.). With *that* as the initial element, demonstrative clefts have a strong anaphoric and attention-marking function (Weinert & Miller 1996: 192f.) and are typically used to underline or sum up previous discourse or to make reference to what has been said before (Collins 1991: 145f.; Weinert & Miller 1996: 192f.; Biber et al. 1999: 961ff.), while those introduced by *this* have a forward-pointing function and are also used as an attention marker (Weinert 1995).

Calude (2008: 99ff.; 108) suggests four discourse functions of demonstrative clefts. For the qualitative analysis of the discourse functions in the present case study, her taxonomy was adopted with slight modifications and two more functions (summarizing and projecting) were added. The six functions are exemplified in turn in (7) – (12).

(7) **quoting**: signaling direct speech, indirect speech or self-reported thought

<B> erm and I I wanted to come to university and do literature <XXX> interested<?> in that .. and it was only really when I was looking through the prospectus sort of thinking well I don't just want to do literature what can I put [with it <\B>

<A> [mhm mhm <\A>

<B> I sort of discovered the linguistics department and thought .. ah yeah **that's what I've always wanted to do** <\B> (LOCNEC)

[11] One may add here that another feature that contributes to their formulaicity is that in contrast to other types of clefts, demonstrative clefts are not reversible (Biber et al. 1999: 961).

(8) **explaining**: giving a reason for a point previously made; explaining how two prior utterances relate to each other (linking function)

<B> yeah I think geography is interesting **that's why I study it** <laughs> </B> (LINDSEI-G)

(9) **evaluating**: giving opinions, evaluations or assessments; expressing agreement, disagreement or a neutral opinion with a previous comment

<B> yeah it wasn't much of a holiday really <\B>
<A> oh no <laughs> <\A>
<B> <laughs> <\B>
<A> it was just a a working holiday <X> <\A>
<B> a working holiday yeah <\B>
<A> just work <\A>
<B> well that's that's <X> **that's exactly what what our bosses were saying** exactly the same phrase said er you're here for no holiday you work you're here to work <\B> (LOCNEC)

(10) **highlighting**: singling out a preceding discourse element, thereby foregrounding it and giving it special prominence

<A> since you like the cinema so much <\A>
<B> [mhm <\B>
<A> [would you like to: to do: .. later to work . in relation . to <\A>
<B> <X> what I'd like to do well I mean my degree is a primary school teaching degree **that's what I'm aiming to do at the[i:] end** <\B> (LOCNEC)

(11) **summarizing**: summing up a longer stretch of previous discourse

<B> he's changed the picture so that she's erm she looks considerably younger .. erm obviously the hair's changed the face has changed <\B>
<A> [mhm <\A>
<B> [she's she's got a slight smile erm .. and then now she's sort of erm just telling all her all of her friends sort of oh this is a picture of me isn't it lovely and doesn't it look so much like me but er \B>
<A> <laughs> <\A>
<B> **that's that's how I would say the story is going** she's er .. she's she's eh this woman is actually quite vain <\B> (LOCNEC)

(12) **projecting**: drawing attention to a following stretch of discourse (only with cataphoric deixis)

<B> so . it was a really nice (erm) . experience . I had and . what I found most (erm) impressive and I think **that's what everybody says when . he has seen Australia** is that . (erm) the distances are so huge . it's (er) that's really amazing so one day we drove for twelve hours and there was nothing . li<?> (eh) it's only dust . around us and so . but . it was really . yes impressive <laughs> </B> (LINDSEI-G)

Previous corpus-based studies of reversed *wh*-clefts in English interlanguage are based on subsets of the ICLE. While Herriman and Boström Aronsson (2009) found an overrepresentation of reversed *wh*-clefts in the writing of Swedish EFL learners when compared to native speaker writing (93 vs. 62 instances), Callies (2009a) noted that native speakers used demonstrative clefts slightly more often when compared to the writing of German EFL learners (27 vs. 19 instances), but this is not statistically significant. Moreover, Callies observed that the learners showed little variation in how they used this construction: *what* was by far the most commonly used *wh*-word in reversed *wh*-clefts by both groups of writers, but the native speakers employed a broader range of *wh*-elements, while *how*, *where*, and *when* were completely absent from the learner data. They also strongly preferred *that* as a deictic marker and used the copula almost exclusively in its contracted form *'s*, which may indicate that the learners saw this as a formulaic expression. Non-deictic elements in reversed *wh*-clefts were exclusively used by native speakers.

In view of these previous research findings and a contrastive analysis of such cleft types in French, German and English (see further below), the following two working hypotheses can be put forward for the case study: 1) demonstrative clefts are underrepresented in both learner corpora when compared to native speaker usage, and 2) advanced learner language is characterized by a narrower range of the formal and functional uses of this construction.

In fact, the quantitative analysis of the frequency of occurrence of demonstrative clefts in the three corpora (Table 2) shows that demonstrative clefts are significantly underrepresented in the L1 French component of the LINDSEI when compared to the LOCNEC (LL= -7.7**), but that there is no statistically significant difference between the LINDSEI-G and the LOCNEC (LL= +0.23).

Corpus	Absolute frequency	Normalized frequency per thousand turns
LINDSEI-F	27	4.72
LINDSEI-G	57	9.42
LOCNEC	73	8.65

Table 2. Frequencies of occurrence of demonstrative clefts in the three corpora

When analyzing the distribution of this cleft type in the two learner corpora, we find a high degree of inter-learner variability. In both corpora, it is merely a handful of learners who provide for almost 50% of all tokens whereas half (or more) of the learners do not use this construction at all.

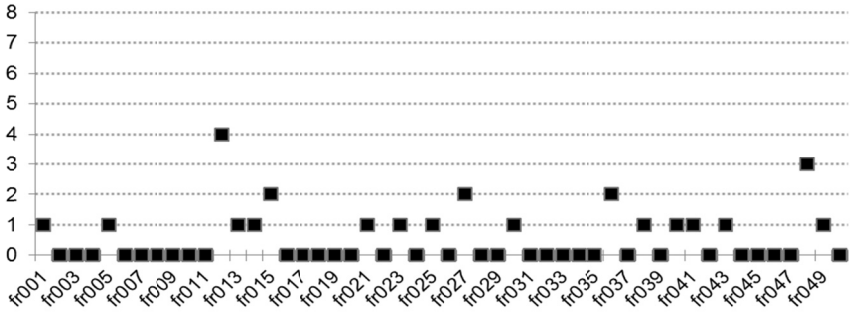

Figure 1. Demonstrative clefts in the LINDSEI-F

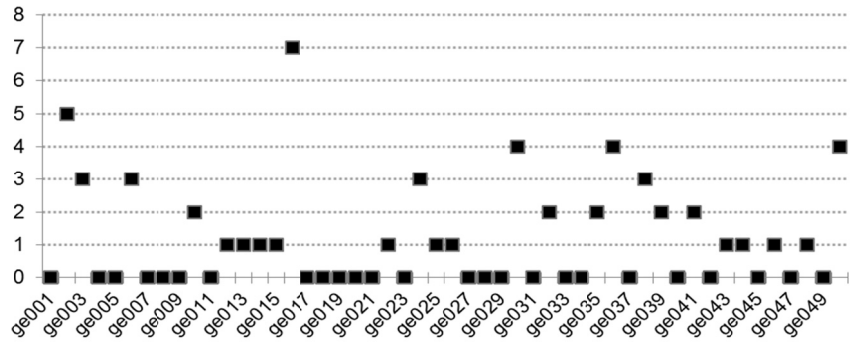

Figure 2. Demonstrative clefts in the LINDSEI-G

The case study thus demonstrates the usefulness of learner corpora to abstract away from individual learners to identify a corpus-based description of a specific learner group while also providing insights into inter-learner variability. The individual differences found for both the French and the German EFL learners have important implications for learner corpus analysis and compilation in that they confirm that global proficiency measures based on external criteria alone are not reliable indicators of proficiency. However, in a substantial part of LCR to date, individual differences often go unnoticed or tend to be disregarded and are thus not reported in favour of (possibly skewed) average frequency counts.

It is interesting to compare the two learner groups and the native speakers as to the relatively fixed structure of demonstrative clefts. Similar to the findings reported in the research literature, the deictic *that* and the *wh*-words *what* and *why* are the most frequently occurring elements (Table 3). Demonstrative clefts primarily convey anaphoric deixis in all three corpora. While it is not surprising that the native speakers employ the full range of options that this construction allows in terms of the use of initial demonstratives, *wh*-words and deictic reference, it is indeed striking to see major differences between the two learner groups. The way how the German learners use this construction very much resembles native speaker usage in terms of structural variation. By contrast, demonstrative clefts are not only significantly underrepresented in the spoken language of French learners, but the degree of formulaicity (or invariability) is also highest in the LINDSEI-F.

	LINDSEI-F	**LINDSEI-G**	**LOCNEC**
demonstrative			
that	26 (96%)	44 (77%)	67 (92%)
this	1 (4%)	13 (23%)	6 (8%)
***wh*-word**			
what	12 (44%)	27 (47%)	30 (41%)
why	14 (52%)	17 (30%)	15 (21%)
where	0	1 (2%)	11 (15%)
when	0	4 (7%)	6 (8%)
how	1 (4%)	8 (14%)	11 (15%)