The potential of purpose-built corpora in the analysis of student academic writing in English

Julia Hüttner

University of Southampton | United Kingdom

Abstract: The trend towards using English as an academic lingua franca has undoubtedly increased the awareness of a need for specific EAP writing instruction and inroads into researching student writing have been made. However, systematic improvements for a theory-informed teaching practice still require more detailed knowledge of the current state of student academic writing, which also takes into account local practices and requirements. Extended genre analysis provides such a means of researching student writing in specific settings. This is an innovative methodology which expands on English for Specific Purposes (ESP) genre analysis (cf. Bhatia, 1993, 2004; Swales, 1990, 2004) to systematically integrate corpus linguistic tools into the analysis and to take into account the special status of student genres. A special advantage of this methodology is that it can be applied easily and successfully to small-scale purpose-built corpora. This paper presents an application of extended genre analysis to a corpus of 55 student paper conclusions produced by non-native speakers in the initial phase of their studies. Findings suggest systematic differences in structure between student and expert genres, as well as a more complex set of differences in lexico-grammar, and especially the use of formulaic language, between research articles and non-native student papers. The implications of these findings as well as of the proposed methodology of corpus-based genre analysis for teaching practice are also discussed.

Keywords: genre analysis, corpus linguistics, student writing, EAP

journal of
WRITING RESEARCH

Hüttner, J. (2010). Purpose-Built Corpora and Student Writing. *Journal of Writing Research, 2* (2), 197-218. http://dx.doi.org/10.17239/jowr-2010.02.02.6

1. Introduction

In addition to its role as prime medium of communication in international academic publications, English as an 'academic lingua franca' is gaining a strong foothold in tertiary education in countries of the expanding circle (Kachru, 1992, p. 356), which traditionally only used their national languages in education. (cf. Graddol, 2006, p. 74; Hyland, 2006, pp. 24-25; Swales, 2004, chapter 2) These typically European non-English-speaking students are now frequently required to read English language materials in their disciplines and to produce at least part of their academic writing in English; thus, mastery of English for Academic Purposes (EAP) is turning into an essential student skill, and no longer the specialisation of language students only.

The requirement of more students to write EAP texts in combination with overall rising numbers of students in Europe should increase the awareness of a need for specific writing instruction, and has done so in English-speaking countries. In Austria, however, as in most other continental European countries, the provision of specialised writing instruction is still limited. Mostly, content lecturers give brief advice on correct ways of quoting sources and language teachers highlight the use of connecting devices, leaving students largely to their own devices in finding out what else constitutes acceptable academic writing. Teaching practice often remains intuitive and the requirements of a good student paper are rarely made explicit.

On a more general level, it seems that systematic improvements for a theory-informed teaching practice still require more detailed knowledge of student academic writing. While the research base on student writing in the L1 is growing and reference corpora of student writing are currently being compiled (cf. the British Academic Written English Corpus, *BAWE,* and the Michigan Corpus of Upper-Level Student Papers, *MICUSP*), the situation regarding research into non-native student writing typical of the situation in continental Europe described above is very different; in general, considerably less attention is paid to student writing in English as a foreign language and corpora on non-native writing such as *ICLE* (*International Corpus of Learner English*) do not include the genres required of students in many European universities. The potential uses of student writing corpora are similar to those described for learner corpora. Thus, they can serve as a basis for research, for material design based on the information regarding problem areas of the student writers, and as a resource for students themselves in discovery learning (data-driven learning).[1] The latter approach can be particularly informative for students if it reflects their own practice and can be compared to some other practice, be it native student or expert writing. The development of larger, non-native corpora of academic English is a desirable long-term goal, especially if modelled along similar lines to *ICLE*, i.e. by including diverse L1s and thus allowing for information on practices of individual language groups as well as of problems encountered by non-native writers in general. One way of addressing this gap more immediately, however, is by creating small-scale purpose-built corpora of student

texts. This contribution will highlight the potential of using such corpora as a basis for analysing student writing with a view towards pedagogical applications.

As will be argued here, the need to research student writing in specific settings calls for a new methodology which takes into account the special status of student writing within academia. The methodology proposed for this aim is 'extended genre analysis' (cf. Hüttner, 2007, 2008), which is firmly embedded in genre analysis in the ESP tradition, following Bhatia (1993, 2004) and Swales (1990, 2004), but developed further to take into account the special status of student genres and to systematically integrate corpus linguistic tools into the analysis. The main innovation of this methodology is to focus on the integrated analysis of patterns of conventionalisation in language at two levels: firstly, the macro-level of the genre structures employed and, secondly, the micro-level of the lexico-grammatical and phraseological profile of the student genres identified. Importantly, the use of small purpose-built corpora allows for including contextual information and acknowledges local norms and practices. As a way of exemplification, this paper presents the findings of an extended genre analysis of a corpus of 55 student paper conclusions produced by non-native speakers in the initial phase of their studies, and addresses the relevance of these findings for teaching.

2. Theoretical Background

2.1 Genre Analysis

The analysis of the writing produced in academic settings has attracted interest from a variety of sub-fields of linguistics, notably genre analysis and corpus linguistics. The arrival of a discourse-oriented genre analysis successfully challenged the previously held view of EAP as a unified whole by showing that the diverse writings produced in academic settings, e.g. exams, reports, essays, papers, lecture notes, etc., follow equally diverse conventions. In the so-called ESP approach to genre analysis (e.g. Bhatia, 1993, 2004; Swales, 1990, 2004), emphasis is placed on the unique sets of communicative purposes that are fulfilled by distinct academic genres. This central position of the criterion of communicative purpose is underlined in the definition of genre as:

> a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style. (Swales, 1990, p. 58)

Arguably the 'master genre' to have been studied in this framework is the research article, and in his insightful analysis of introductions, Swales (1990, pp. 137-166; 2004, pp. 226-234) exemplifies how authors include important persuasive and 'territorial' purposes in trying to "create a research space". They do this by showing the value of

their research through making it rhetorically fill a gap left open by previous research. (cf. also Anthony, 1999; Kwan, 1996; Lewin, Fine & Young, 2001; Nwogu, 1990; Samraj, 2002) Moreover, connections between the choices made on a purely linguistic level to the communicative purposes of the authors have been established in this framework, pointing out also minute differences in disciplinary conventions. (cf. e.g. Anthony, 1999; Ozturk, 2007; Samraj, 2002; 2005) In sum, genre analysis created a new understanding of EAP as consisting of a variety of individual genres, partly clustered in 'genre-colonies' (Bhatia, 2004, pp. 57-58) related by similarities in purpose or by disciplinary affiliation.

One of the criticisms that can be raised against genre analysis, in line with many other forms of discourse analysis, has been its reliance on relatively small data bases. Lee (2008, p. 88) notes that "for some, discourse analysis can proceed very well [...] with just one text, and manual (painstaking) analysis is almost *de rigueur* for some analysts". While Lee considers the reason for such hesitance towards corpus-based methodologies in the more technical aspects involved, including the accessibility of corpora, a further possible reason might be the assumption that any individual text can serve as a proto-typical example of the genre it is part of. This is problematic as extended genre analysis indicates quite clearly that one important aspect of any analysis is finding out what exactly constitutes a proto-typical genre exemplar that is accepted by the discourse community of genre users. In order to objectify intuitions and to identify move and language structures that are indeed typical of any specific genre, information from a larger number of texts has to be abstracted.

2.2 Corpus linguistics

One means of addressing this issue is to take recourse to corpus linguistic methods, and extended genre analysis presents one way of doing so. In this combination, extended genre analysis takes up on earlier calls propagating the mutual benefit of discourse studies and corpus linguistics (cf., e.g., Hardt-Mautner, 1995; Kirk, 1996, p. 276), which have so far not led to a systematic integration of the two approaches, despite some impressive attempts (cf., e.g., Tribble & Scott, 2006).

The reason why linguistic corpora, i.e. electronic compilations of texts, are helpful in establishing typical patterns of language use lies in the fact such patterns have been shown to frequently escape intuitions of native speakers and of teachers. One example of this is that corpora can reveal that speakers favour particular lexical choices or combinations over (near-synonymous) alternatives, but native speakers are rarely able to make this procedural language knowledge explicit. Corpus linguistic methods offer such means of making these patterns apparent through a fast and comparatively easy analysis of considerable numbers of texts with the help of specialised computer software.

Large-scale corpora aim to reflect general language use, and so typically consist of both written and spoken texts.[2] Findings from such corpora have provided information on issues such as the typical choice of words (i.e. their frequencies), the collocations

that words typically enter into, including the formulaic sequences or multi-word chunks that recur in language use. Additionally, the 'semantic prosody' of words, i.e. the meaning nuances associated with near-synonyms, can be established with the help of corpora. Examples of such semantic prosodies include the typical combinations of the word *cause* with negatively connotated items, such as *anxiety, problems, cancer or damage* (cf. Stubbs, 1996, pp. 176-181).

Corpus-derived information such as this has fed into the compilation of dictionaries, especially those geared to learners (e.g. MacMillan, Cambridge) and teaching materials. However, despite the possibilities and applications of these corpora, one of their major limitations concerns their representativeness, i.e. the question of how well any corpus, regardless of its size, can reflect 'a language' in its totality. (cf. Hunston, 1995; Kaltenböck & Mehlmauer-Larcher, 2005; McEnery & Wilson, 1996) Thus, some genres tend to be under-represented in general corpora while others are over-represented, which can lead to the absence of particular genres and the terms associated with them. A case in point is the term *letter of credit*, an item typically required in business English, which is absent from such a comparatively large corpus as *ICE-GB*. Especially for researchers and teachers of languages for specific purposes, such gaps can be crucial. After all, if a particular teaching context involves students having to learn how to write lab reports in the field of biology, a corpus that includes many journalistic texts and even some biology textbooks or abstracts of biology dissertations, but no such lab reports, will not provide the information needed.

This situation provides reasons for using purpose-built corpora of specific genres; thus, in the situation above, a teacher might benefit from the information drawn from a small corpus of only biology lab reports, which will give information on the particular use of language in biology lab reports. Such corpora allow for the focused study of genres that are rare in general language use, that have just emerged, or that are occluded, i.e. not accessible to the general public, such as educational genres.

2.3 Studying student writing

Investigations into student writing have begun to receive attention from diverse directions, often in direct association with rising student numbers and perceived increased needs for specific instructions in writing within academia. One aspect that becomes clear soon, however, is that despite this interest and the inroads made with regard to native student writing, our knowledge of student writing in English in non-English-speaking countries still needs to be expanded before instructors have more than their intuitions to base their teaching and marking on.

This knowledge relates to two important aspects and neither of these is exclusively limited to non-native students. Firstly, the structures employed by student writers frequently differ from those considered appropriate by their markers. A clearer awareness of these differences could lead to two possible actions. The first would be fairly straight-forward in the recognition that many student writers benefit from being given clear and transparent guidelines on the structures expected of their writing and to

make these available to students. As this is not done in all relevant educational contexts, including the Austrian one, students are left to find their own models, either resorting to school writing models, which is generally penalised, or to expert models, which have quite diverse communicative purposes and so are often rather inappropriate for student writing. Frequently such expert models are suggested to students either explicitly (cf. e.g. Swales & Feak, 1994, pp. 155ff) or implicitly by assuming that students would find a suitable model for their papers in their academic language environment.[3] As textbooks and lectures are probably the genres most frequently encountered by students this, however, leads to problematic genre mixtures at the structural level. A further and arguably more innovative reaction to such an awareness would be for teachers to recognise potentially valid communicative intentions of student writers and thus allow some negotiation of what should be included in student academic writing or to allow for different forms of outlet for these communicative needs. The second area where information on student writing is beneficial to teaching practice involves more precise knowledge of lexico-grammatical and phraseological patterns employed by students vis-à-vis experts.

3. The rationale and methodology of extended genre analysis

The methodology for the analysis of student genres proposed here, i.e. extended genre analysis, addresses several problems mentioned earlier. Firstly, it aims to take into account the special situation of student writing, where an adequate analysis of the genre has to take into account not only the texts students are producing as samples of a particular genre (i.e. the students' role as 'genre owners'), but also the role of lecturers as markers of the genre exemplars produced by students and so as gate-keepers or 'secondary' genre owners. Thus, it gives information on the typical structural patterns in student writing and their evaluation by experts. It so allows for the identification of 'sites for negotiation', i.e. structural elements typically employed by students, but considered inappropriate by markers.

Secondly, this methodology systematically integrates corpus-linguistic methods into a genre analysis as part of the linguistic analysis. In this way, the advantages of genre analytic methods are enriched by corpus-linguistic methods that enable detailed comparisons with both patterns of general language use and with patterns of language use in related genres. Special attention is paid to formulaic language use, i.e. to "sequence[s] [...] of words or other meaning elements, which [are ...] prefabricated" (Wray, 1999, p. 214) and so this methodology acknowledges co-conventionalisation, i.e. the existence of conventionalisation in language use on both a macro- and a micro-level of genre and formulaic language use respectively. (cf. Hüttner, 2007; 2008)

Indications of such co-conventionalisation in previous research largely take the form of the observation that some formulaic sequences are typical of specific speech events or genres in that they occur predominantly in these. This correlation has been shown to occur in native and learner language use, and has been established for a

variety of genres or speech events, such as auctions, check-out transactions at the supermarket, weather forecasts, lectures, academic texts and children's games. (cf., e.g., Biber, 2006; Jones & Haywood, 2004; Kuiper, 1991; Kuiper & Flindall, 2000; Linnakylä, 1980; Schmidt, 1983; Wong Fillmore, 1979)

Phase 0: *Corpus Selection*

- Random selection of texts considered exemplars of genre X by discourse community or selection of entire cohort of texts considered exemplars of genre X. For student genres: all texts submitted as instances of genre X

Phase 1: *Linguistic Analysis of Corpus – describing the status quo*

- Analysis of move structure
  - preliminary decision on status of individual moves as obligatory, core or optional on quantitative basis
  - preliminary decision on proto-typical sequences on quantitative basis
- Linguistic analysis
  - lexico-grammatical analysis:
    - describe lexico-phraseological profile of genre (keywords, key phrases)
    - make considered use of comparative corpora (large reference corpora, corpora of related genres, etc.)
    - describe typical grammatical patterns overall and for individual moves
  - patterns of textualisation
    - describe typical patterns of textualisation of specific communicative intentions

Phase 2: *Systematic integration of expert information – arriving at a model for teaching*

- choose expert informants familiar with the institutional context and the genres studied
- present a sub-section of texts to appropriate expert members for detailed comments on appropriacy and acceptability
- present examples of individual moves to informants for evaluations regarding acceptability/appropriacy
- use this information to come to a final decision on the status of individual moves and sequences of the genre structure developed in *Phase 1*

Figure 1. Extended genre analysis: Overview of methodology.

In the analytical framework of extended genre analysis outlined below, these observations are taken onto another level by explicitly linking formulaic sequences to genre. Two types of formulaic sequences are proposed in connection with genre; firstly, *key formulae*, i.e. recurring multi-word chunks that are quantitatively typical of

particular genres and thus constitute part of their specific 'idiomaticity', and secondly, *genre-functional formulae*, i.e. those sequences that further the communicative purposes of a particular genre move (cf. Hüttner, 2007, pp. 97ff).

The details of the proposed methodology of extended genre analysis are presented below.

Phase 0 incorporates one of the most important alterations to traditional genre analysis by abandoning the notion of a pre-selection where only proto-typical texts are included in the selected corpus. Instead, all members of a particular genre are considered the target population here and members of the corpus are selected – as far as possible - randomly for inclusion, allowing for the discovery of frequent genre moves that might not feature in examples pre-defined as typical and possibly therefore as 'good' examples.

The linguistic analysis takes place in two phases, starting out with Phase 1 where the analysis of the corpus selected in Phase 0 occurs to be refined through the systematic integration of expert/gatekeeper information in Phase 2. One of the tasks in the linguistic analysis of Phase 1 is to describe the move structure of the genre. Moves in this framework signal functional parts with specific communicative intentions which together constitute the overall communicative purpose of the genre. More precisely, a move is "a discoursal or rhetorical unit that performs a coherent communicative function in a written or spoken discourse" (Swales, 2004, p. 228). This level of analysis essentially remains to be conducted by the researcher (or, ideally, the team of researchers) manually as the primary decision is based on rhetorical purpose. Thus, the moves need to be defined through researchers' careful analysis of the entire texts with an aim of capturing all functional elements, and by taking into account some linguistic cues in the decision of move boundaries. (cf., e.g., Connor & Mauranen, 1999, p. 50 for a detailed description of this process) Although some contextual knowledge is required, this process is descriptive rather than prescriptive. It does not set out to find particular moves, but tries to identify stretches of text that fulfil a function recognisable to users of the genre. In order to increase the reliability of such move decisions, work in teams of researchers is ideal, or, as a minimum requirement, checking with other researchers on move decisions. Although some practice is required for such move-analysis, it is remarkably quickly learnt. (cf. Hüttner *et al.*, 2009 on the applicability of this approach to student teachers)

In Phase 1, the focus lies on finding quantitatively based typicalities both on a lexico-grammatical/phraseological level and on a level of genre structures, i.e. definition and order of moves, by taking all examplars of the corpus into consideration. With regard to the genre structures, preliminary decisions on the status of individual moves as obligatory, core or optional within the genre are here based on their frequency of occurrence. I propose the following preliminary quantitative measures:

Table 1. Guidelines for deciding on status of individual moves

Frequency of Occurrence	Status	Comments
90% - 100%	obligatory	genre exemplar usually considered inappropriate or in some way "flawed" *without* this move
50%-89%	core	typical of the genre, considered part of an appropriate and acceptable genre exemplar
30% - 49%	ambiguous	status can only be decided with further expert information – can be core or optional, acceptable or unacceptable (Phase 2 decisive)
1% - 29%	optional	not considered a typical feature of genre, can be considered an acceptable addition (=truly optional) move or inacceptable (-> Phase 2 decisive)

As shown in Table 1, the status given to individual moves in Phase 1 ranges from obligatory to optional, and a 'model' genre structure is expected to consist of only core and obligatory moves. We can see here already that the frequency bands below 50% rely on further information from experts regarding their status and, importantly, their acceptability. This takes account of the fact that a move in a student genre can be considered unacceptable by the gatekeepers despite comparatively high levels of occurrence

Phase 2, i.e. the systematic integration of gatekeeper information, is particularly relevant as student texts are subject to marking procedures and not all find acceptance from the relevant gatekeepers, i.e. their lecturers. It is more qualitative in nature and involves obtaining information from expert members of the discourse community on the acceptability and appropriateness of individual moves as well as of entire texts. These expert judges have to be experienced members of the discourse community under investigation so that they know what is acceptable. The information obtained in this way provides confirmation or refutation of the status of all moves as obligatory, core or optional. What is even more important is the information gained on which moves might be considered inappropriate or unacceptable. In this sense, Phase 2 takes the expert view on the student contributions and thus shows what of the students' views – encoded in their submissions analysed in Phase 1 – is supported by the gatekeepers of the discourse community. After these two phases, a model of the core move structure can be arrived at which reflects both students' actual performance and the acceptability of this model from the perspective of the markers.

The benefits of having arrived at such a model for student genres are, firstly, that it can easily be made transparent to students, who are then no longer left to 'find out for themselves'. Secondly, teachers can benefit from making their intuitive knowledge explicit. In this context, the reaction of one of the expert judges of the study described below was revealing as she said that being forced to comment explicitly on the appropriacy of individual texts made her realise more clearly than before what were to her the most important elements in student conclusions. Finally, the availability of such

a transparent writing model raises awareness for the potential of changing it, possibly by taking on board some of the moves reflected in student writing and not currently accepted by (all) markers.

Further levels of linguistic analysis focus on lexico-grammar. This incorporates corpus linguistic methods by using a variety of reference corpora to highlight genre characteristics, and a close analysis of textualisations of individual moves in order to highlight the means employed to advance specific communicative intentions. In this step, the potential of using corpus-linguistic tools such as WordSmith Tools (© Mike Scott) to develop information on the lexico-phraseological profile of a particular genre is highest. In order to establish a lexical profile of a particular genre, the first step is to create a so-called *keyword* list (cf. Tribble & Scott 2006, chapter 4). This provides the researcher with a list of lexical items that are statistically typical of a particular corpus, either by occurring more frequently *(positive keywords)* in this specific corpus or less frequently *(negative keywords)* than in a general language corpus. The statistical comparison is achieved by first creating a word-list of the genre under investigation, i.e. a list of all the words and their occurrences in the corpus. This wordlist is then compared with the wordlist of another, larger corpus, the so-called reference corpus. This can be a large corpus of general language use, such as the *BNC World*, which will then give a rough comparison with general English language use, or a more closely related corpus, e.g. in our case of investigating student writing, a corpus of expert academic writing in the relevant discipline. Keyword lists give clues as to the lexical profile of the genre under scrutiny; such information can be valuable in teaching contexts as it indicates which words are required in the production of these genres.

The way in which keywords are employed, i.e. the collocations they enter into, are of interest in establishing a phraseological profile of a genre; here corpus tools support the researcher in several ways, firstly, by establishing so-called *concordances*, i.e. retrieving the immediate co-text surrounding the keyword. What emerge are the typical patterns, be they lexical, syntactic or 'semantic prosodies'. Additionally, WordSmith Tools enables researchers to find recurring clusters, i.e. multi-word units, from within the entire corpus.

With respect to formulaic language use, as mentioned earlier, two types are proposed in connection with genre; firstly, *key formulae*, i.e. recurring sequences that are quantitatively typical of specific genres. An example of such a *key formula* would be a chunk like *a large number* or *the development of*. The second group proposed are *genre-functional formulae*, i.e. sequences that further the communicative purposes of a particular genre move. (cf. Table 2)

Table 2. Comparison genre-specific vs. genre-functional formulae

key formulae	**genre-functional formulae**
linked to overall genre similar to lexical key-words defined as typical by frequency of occurrence within genre sub-divisions possible, e.g. ▪ across a range of technical to non-technical ▪ grammatical patterns ▪ genre-referential items	linked to specific genre moves defined by frequency **and** by furthering the communicative purpose of the genre one feature of move-textualisation

Depending on the genre, and on individual moves, the dominance of formulaic sequences or particular types of sequences might vary. In contracts of sale, for instance, a preliminary study (Kastenberger, 2005) found extensive use of genre-functional formulae, which highlighted the communicative intentions of nearly every move. One might argue that producers of especially this legal genre of contracts need to ensure that all important communicative intentions of the genre text are fulfilled and are textualized in completely unambiguous ways. In fact, the actual formulaic sequences might not even be readily understood by all users, but serve to indicate which pieces of information are about to come, and in a way act as a frame for the important content of the precise goods to be sold, prices arranged, and obligations entered into.

In other genres, less use might be made of genre-functional formulae or this use might be restricted to specific moves. Especially in emerging genres or genres where creative language use is highly valued, formulaic sequences are in general likely to occur less frequently. Regardless of the actual frequencies of occurrence, however, investigating the link between formulaic sequences and genres systematically is a vital improvement for both fields of research; on the one hand, genre studies are furthered by considering this aspect of conventionalised language as a logical extension of studying conventionalisation at a structural level. On the other hand, the study of formulaic language use is given a clearer focus by limiting its investigation to one specific genre at a time and systematically differentiating between formulaic sequences that are functional in a specific language context, and those that are not, but might still be a defining part of the phraseological make-up of a genre.

4. Student conclusions – Findings from extended genre analysis

The empirical study presented here is an application of the above methodology. The data base under investigation here comprises 55 student academic papers on linguistics written by students of English at the University of Vienna. All students were native speakers of German and for the majority (77.2%) these papers constitute their first academic papers in English. These papers form part of the course requirements of an

introductory class in linguistics, are written in English and constitute an in-depth treatment of a particular topic in one of the fields of the courses (e.g. pragmatics, mental lexicon, sociolinguistics, grammar) of about 3,000 words in length. The entire corpus of student papers amounted to about 160,000 words. Some of the papers are based on student projects, while others are library-based only. They do not correspond to British-style student essays, but are more comparable to project papers in British or American settings or to Seminar papers in other European settings. The writing task itself was not that clearly described in the classes studied, where usually students picked a topic and presented on it first orally and then wrote their papers on it. Most lecturers gave some feedback after the oral presentation, in some cases making explicit reference to what should be different in the written paper. Little time, however, was devoted to providing information on academic writing. The instruction offered focused mostly on the appropriate use of sources and the guidelines for quoting and referencing. All students had experience of reading textbooks in linguistics in English, and some articles were required reading in the courses. (cf. Hüttner, 2007, pp. 120-125)

In the following, I will present an analysis of the *conclusions* of the student papers investigated. The corpus of student conclusions consisted of 55 texts of quite diverse length, amounting to 10,861 words in total, with a mean of 197.5 words (std.dev. 113.09). This section in the paper is perceived as rather problematic by students, seeing that it is inherently dense, requires some synthesis of what has been previously written and does not allow the student authors much reliance on secondary sources – quite in contrast to the body of the paper. Comparative data was elicited through an analysis of 55 expert research article conclusions from the field of linguistics, amounting to a total of 44,523 words, with a mean length of 809.05 words (std dev. 619.38). The articles for the expert corpus were randomly chosen from 13 different international edited journals that deal with similar topics to those covered by the student writers. The authors were assumed to be either native speakers of English or at least highly competent L2 users of English from surnames and professional affiliation. (for a complete list, see Hüttner, 2007, Appendix)

4.1 Move structure

The following table presents the move structure identified, with the moves shaded in grey representing the core moves of the genre, i.e. those that were confirmed by both quantitative and qualitative investigation, i.e. phases 1 and 2 of the methodology.

Table 3. Frequency of move realisations in student paper conclusions

Move	Total (N=55)	% Total
Provide a Summary Statement or Review	53	96.4%
Qualify and Evaluate the Paper/Results	32	58.2%
Provide a Personal Reflection	18	32.7%
Provide a Wider Outlook/Embedding Paper	18	32.7%
Present New Information	13	23.6%
Appeal to Reader	9	16.4%
Acknowledging Gratitude	2	3.6%

This move structure quite clearly differs from the one that could be established for expert research article conclusions (cf. Hüttner, 2007; Lewin *et al.,* 2001). Noticeable differences lie in the fact that expert conclusions have a fully obligatory move of REPORT ACCOMPLISHMENTS to focus on the results; in student conclusions this was combined with the REVIEW move. The latter was realized in some expert articles, but only in *addition* to the obligatory REPORT ACCOMPLISHMENTS move. While this might seem rather minimal, it does show that even if it is acceptable for students to simply run through what has been presented in the paper again, for experts it is vital to draw a conclusion. The importance for experts of placing their research and, in a way, of defending their future research space, can be seen in the frequency of occurrence of the STATE IMPLICATIONS move. Indeed, the future research outlined in that move frequently corresponds to the author's research agenda, and for this reason arguably would need to be 'marked' as his or her ideas as soon as possible. In the case of student papers, however, mentioning future research possibilities tends to be quite a vague acknowledgement that more work could be done in a particular area. The most important differences lie the experts attaching importance to warding off any potential criticism on their research by emphasizing the value of their results despite – minor – limitations in the WARD OFF COUNTERCLAIMS move (cf. Lewin *et al.,* 2001, pp. 65ff).

I would argue that this move structure, which is shown to be both achievable by students, seeing that it reflects their actual contributions, and endorsed by the institutional gatekeepers should be the one to be used as a model in teaching this genre. It is worth noting that four moves were excluded from the final move structure following Phase 2 of the methodology; reactions to these moves were quite different, with APPEAL TO READER and ACKNOWLEDGING GRATITUDE were considered more optional and possibly slightly humorous but PRESENT NEW INFORMATION and PROVIDE A PERSONAL REFLECTION were considered quite seriously unacceptable. (cf. Hüttner, 2008 for more details on the development of this move structure and especially on the status of the REFLECTION move)

4.2 Lexical profile

The corpus-based lexico-grammatical analysis of this genre focused on two aspects; firstly, on the development of a lexical profile of the entire genre by establishing keywords. These suggest quite clearly that despite evidence of a growing familiarity of

the authors with the disciplinary terminology shown in their use of technical terms, the high ratio of non- and semi-technical vocabulary indicates that this type of student writing is still situated somewhat in between personal or school genres and the more discipline–specific, technical genres that students are becoming familiarised with at university. This picture differs from expert writing, where the cline is more towards the technical terms. This can be seen as possible evidence for the status of the student authors as peripheral members of the discourse community involved. Table 4 shows the technical and non-technical items established in the top 100 keywords for students and experts, leaving aside the largest group of semi-technical items. Although experts and students both use technical as well as non-technical items, this comparison makes it quite apparent that the ratio is different.

Table 4. Experts' vs. students' use of technical vs. non-technical keywords

Experts		**Students**	
Technical	**Non-technical**	**Technical**	**Non-technical**
acquisition	apes	acquisition	advertisement
aphasic	differences	anglicisms	advertisements
aphasics	different	borrowings	advertisers
constraints	effect	cohort	advertising
coreference	English	ellipsis	are
CS	gossip	hypothesis	Austria
CSs	humor	lexicon	Austrian
elaboration	immigrant	linguistic	Austrians
excitation	in	maxim	behavior
fricative	joking	maxims	callers
glottal	of	overextensions	children
integrativeness	present	prototype	confetti
lexical	self	prototypes	differences
linguistic	signalling		different
multilingualism	students		diversity
NESB	suggest		English
nonfluent	these		found
nouns	this		frequently
reflexive	underlining		gender
semantic	use		general
semilingualism	Welsh		German
socio			have

sociocultural	humour
vocabulary	important
voicing	interesting
	jokes
	more
	opinion
	our
	people
	sexist
	show
	shown
	taking
	that
	this
	use
	violation
	we
	women

One interesting aspect in classifying the items on the keyword list is the fact that cutting across the division into technical, semi-technical and non-technical vocabulary, we find a group of keywords relating to the genre itself. These are terms that are not content-related as such in that they do not focus on the various topics of the student conclusions. This group draws its members from all three groups of the technical to non-technical categorisations, showing that this is not just an additional category but a diverse way of categorisation. We find here lexical items addressing diverse parts of the student paper conclusions or the student paper, such as *conclusion* and *paper,* a cluster of items referring to research activity, i.e. *data*, *interviews*, *questionnaires*, *research*, some general items like *addressed*, *mentioned*, *investigated*, *topic* and finally a group that refers to the presentation of conclusions and review of results, i.e. the most important communicative purpose of the genre-constituent. Terms include *conclude, conclusions, consequently, proved, results, show* and *summing*.

Table 5. Genre- referential keywords (student paper conclusions: reference corpus BNC)

addressed	cf.	conclude
conclusion	conclusions	consequently
Data	hypothesis	important
interviews	investigated	mentioned
opinion	paper	proved
questionnaires	research	results
show	shown	summing
topic		

There is also one interesting negative keyword, namely *study*, which is typically absent from student texts when compared to expert research article conclusions. This is, I believe, an indication of the fact that the term *study* is one of the typical genre-referential expert items, which is contrasted with such words as *research* and *paper* in the student counterparts. A contributing factor to the avoidance of this term could be that for students *study* has connotations of learning rather than of research projects.

The group is clearly prominent within the entire list of keywords (amounting to about 20% of all keywords) and supports the notion that – at least student genres – are made up of two major groups of keywords, i.e. those that highlight the topic or discipline and those that refer back to the genre itself. Some of these keywords occur in both key formulae and genre-functional formulae, discussed below, arguably providing supportive evidence of student authors' greater need to be explicit in formulating their writing and communicative purposes in their texts.

4.3 Phraseological profile

With respect to the two types of formulaic sequences postulated in this framework, we can firstly note that both types, i.e. key formulae and genre-functional formulae, feature in student paper conclusions.

4.3.1. Key formulae

Key formulae, i.e. those sequences that are typical of the genre in question in a similar way to lexical keywords, constitute the larger group and can be sub-divided into several groups, most importantly, discipline-based and genre-referential. The discipline-based key formulae show the familiarity on the part of the students with the phraseological typicalites of their discipline's technical and semi-technical keywords, including typical collocations and some grammatical patterns related keywords, such as <Noun> *of* constructions in chunks like *aspects of* or *analysis of*. The examples observed are given in the table 6:

Table 6. Student key formulae (discipline based)

Aspects of	L2 vocabulary
analysis of (the)	language acquisition
English borrowings	language learners
English expressions	language use
face saving	maxim of (quantity)
face threatening	mental lexicon
face to face conversation	native speaker(s)
first language	non-verbal communication
foreign language	non-verbal behaviour/or
foreign language learners	non-verbal signals
gender-specific	prototype theory
German English borrowings	turn taking (in)
intercultural communication	

The second group, i.e. genre-referential key formulae, consists of sequences that serve to either refer to and/or comment on research activities conducted, the student paper itself or to help structure the text for the reader.

Table 7. Student key formulae (genre-referential)

been proved	our data
can say that	our paper
data has	our research
found out that the	our research paper
good data	paper I have
has shown	shown in
I hope that	shown that
I think that	the conclusion that
I would have	the opinion that
if we had	this research paper
important role	very important
in general	very interesting
in our minds	we can say that
in this paper	we have seen
my hypothesis	we want to
my research	would have been
of the opinion that	we have seen

It is worth noting here that while the same groups could be observed also in expert conclusions, the ratio was again markedly different; while in student conclusions the numbers of genre-referential sequences outnumber the discipline based sequences, the reverse is true for expert research article conclusions, which additionally show a decidedly larger number of discipline-based formulaic sequences. The precise numbers are given in table 8 below:

Table 8. Discipline-based vs. genre-referential key formulae in expert and student conclusions

	discipline-based	**genre-referential**
experts	116	39
students	24	34

4.3.2. Genre-functional formulae

The second type of formulaic sequences established in this framework consists of genre-functional formulae, i.e. those sequences linked to particular moves within the genre that further the communicative intention of that move. In the SUMMARY STATEMENT/REVIEW move of the students, the following chunks were observed:

Table 9. Student genre-functional formulae (Summary Statement/Review move)

summing up	[this] proved to be
as a conclusion	we found out that
to conclude	in [this/ my, research] paper I have

In this group, with respect to the items on the left in bold, some clear similarities could be observed with expert genre-functional sequences as realised in the REPORT ACCOMPLISHMENTS move, i.e. *in conclusion* and *to conclude*. As regards the other genre-functional formulae, one aspect that is quite apparent in the students' realizations is a marked absence of hedging devices. This is also supported by some genre-referential key formulae, like *been proved* or *we want to* as well as examples observed in this move, like:

It is obvious that shortly the facts about [x] are:
In the course of my research I found out
my hypothesis has been proved right
the analyses have shown that the working hypothesis of this paper can be proved:
all in all we therefore see our hypothesis, namely X, [proved]
we can confirm that
the examples chosen show this clearly

That such hedging is, however, typical of expert conclusions is indicated in the use of the genre-functional sequences incorporating the term *suggest* in realisations of the expert move OFFER INTERPRETATION (cf. table 10 below).

Table 10. Expert genre-functional formulae (offer interpretation move)

the results of this study suggest that
the present results suggest that
Findings in this study suggest that
this would suggest that

These sequences combining the terms *suggest* with the terms *study*, *research* or *results* are evidence of hedging being employed by experts and are neatly contrasted with the use of clusters surrounding keywords such as *prove, show* and *find out* used by student writers to report on their results. In these phraseological patterns, we can find support for the notion that hedging is problematic for student writers of academic texts. That this is not primarily an issue of L1 influence could be attested in Hüttner (2007), who found that native speaker students produced the same absence of hedging in their writing. Thus, while student writers use rather general concluding formulaic sequences, which might also be familiar to them from other types of conclusions, the experts have a group of genre-functional formulae which are related more clearly to expert conclusions only.

In the student move of QUALIFYING AND EVALUATING THE PAPER/RESULTS, where frequently the limitations of the present paper were acknowledged, the genre-functional formula *X is/goes beyond the scope of this paper* occurred.[4] Arguably, this sequence offers students the possibility of expressing the communicative purpose of acknowledging limitations in a suitable way.

5. Implications for teaching practice

This paper has introduced extended genre analysis as a novel approach towards the study of student academic writing. The analysis of the data presented here has shown systematic differences between student and expert academic writing, both at a structural and at a phraseological level. I would argue that this methodology can be applied effectively by teams of researchers and teachers for local contexts and that the findings of such analyses can inform teaching practice by providing teachers with more objective information about their students' writing practices. Additionally, corpora created for this research can be made available as resources for students' own corpus-based 'discovery-learning'. (cf. Bernardini, 2000, p. 227; Kaltenböck & Mehlmauer-Larcher, 2005, pp. 78ff)

The findings presented here have been made available to teaching practice in a number of ways. Firstly, they fed into a collaborative project to make information on conducting and presenting research available on an e-learning platform to all students of courses with academic writing requirements at the English Department of the University of Vienna.[5] The information drawn from this research included the provision of transparent information on required move structures in introductions and conclusions. This included the abstract guidelines of which information to include as well as examples from student texts.

In my own classes, I focused on the differences in the genre structure present in the student texts and the genre structure accepted by the gatekeepers. Thus, I explained in greater detail what kind of information should be included in the conclusion, and why the PRESENTING NEW INFORMATION move was inappropriate. I felt, however, that the REFLECTION move needed a different treatment as it seemed to be evidence of a genuine

communicative need of the student authors that should not be just brushed aside. Therefore, I acknowledged this need by providing an outlet for this in a different, voluntary, assignment of a short personal reflection to be handed in separately from the paper. This accepted students' need as genuine, while taking into account the reactions of expert judges that this information should not be part of the academic papers submitted. This reflective assignment also provided interesting information on the positive and negative emotions experienced by first-time student researchers and academic writers.

In my classes, I made the use of hedging in expert writing and the lack of it in student writing a topic for discovery-learning by using the available corpora. Students were given the task of checking for particular keywords in both corpora, namely *want*, *would*, *suggest, show* and *prove* and to make notes on their occurrence. This helped raise awareness of this feature of expert academic writing and was followed by discussion of the reasons for hedging and some practice in re-formulating unhedged student examples. These tasks did not take up much class time as the students were familiar with simple corpus linguistic searches. Informal feedback of students indicated that they both enjoyed the task and felt it was motivating to look at 'real' examples that other students had produced.

Arguably the most important effect of this research in my local context was that it gave rise to more discussion among the linguistic section of the Department of English on the writing provisions made for students and the assessment practices employed. Such processes are, I believe, essential for applied linguistic research to make the link between research and practice work ultimately towards a better, theory- informed teaching practice.

Notes

1. See Kaltenböck & Mehlmauer-Larcher, 2005, for an overview of the potentials of computer corpora in language teaching.

See http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/CBLLinks.htm for an overview of corpora in English.

See Gruber, Reisigl, Muntigl, Rheindorf, Wetschanow, & Czinglar, 2006, section 3.2.3, on the effects of inadequate expert models.

Although this phrase did not occur in expert conclusions two occurrences of this genre-functional formula were observed in expert introductions to research articles (cf. Hüttner, 2007). This possibly indicates that this is also an expert formula for acknowledging limitations, even though it seems linked to a different move in a different section of the expert research article.

This project was co-ordinated by Angelika Rieder-Bünemann. Collaborators were Gunther Kaltenböck, Karin Lach, Ute Smit and the author.

References

Anthony, L. (1999). Writing research article introductions in software engineering: how accurate is the standard model? *IEEE Transactions of Professional Communication, 42*, 38-46. doi: 10.1109/47.749366

Bernardini, S. (2000). Systematicising serendipity: proposals for concordancing large corpora with language learners. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 225-234). Frankfurt a. M.: Peter Lang.

Bhatia, V. K. (1993). *Analysing genre: language use in professional settings*. Harlow: Pearson Education.

Bhatia, V. K. (2004). *Worlds of written discourse: a genre-based view*. London, New York: Continuum.

Biber, D. (2006). *University language: a corpus-based study of spoken and written registers*. Amsterdam & Philadelphia: John Benjamins. doi: 10.1075/scl.23

Connor, U., & Mauranen, A. (1999). Linguistic analysis of grant proposals: European Union research grants. *English for Specific Purposes, 18*(1), 47-62. doi: 10.1016/S0889-4906(97)00026-4

Graddol, D. (2006). *English next: why global English may mean the end of English as a Foreign Language*. The British Council.

Gruber, H., Reisigl, M., Muntigl, P., Rheindorf, M., Wetschanow, K., & Czinglar, C. (2006). *Genre, Habitus und wissenschaftliches Schreiben*. Münster: LIT Verlag.

Hardt-Mautner, G. (1995). 'Only connect': critical discourse analysis and corpus linguistics. *UCREL Technical Papers, University of Lancaster (Department of Linguistics) 6*.

Hunston, S. (1995). Grammar in teacher education: the role of a corpus. *Language Awareness, 4*(1), 15-31. doi: 10.1080/09658416.1995.9959864

Hüttner, J. I. (2007) *Academic writing in a foreign language: an extended genre analysis of student texts*. Frankfurt a. M.: Peter Lang.

Hüttner, J.I. (2008). The genre(s) of student writing: developing writing models. *International Journal of Applied Linguistics, 18*(2), 146-165. doi: 10.1111/j.1473-4192.2008.00200.x

Hüttner, J.I., Smit, U. & Mehlmauer-Larcher, B. (2009). ESP teacher education at the interface of theory and practice: introducing a model of mediated corpus-based genre analysis. *System, 37*(1), 99-109. doi: 10.1016/j.system.2008.06.003

Hyland, K. (2006). *English for Academic Purposes: an advanced resource book*. London & New York: Routledge.

Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sequences. In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 269-300). Amsterdam & Philadelphia: John Benjamins. doi: 10.1075/lllt.9.14jon

Kachru, B. (1992). Teaching World Englishes. In B. Kachru (Ed.), *The other tongue* (pp. 355-366). Urbana & Chicago: University of Illinois Press.

Kaltenböck, G., & Mehlmauer-Larcher, B. (2005). Computer corpora and the language classroom: potential and limitations of computer corpora in language teaching. *ReCALL, 17*(1), 65-84. doi: 10.1017/S0958344005000613

Kastenberger, A. (2005). Genre analysis: Contracts of sale. Unpublished project paper. University of Vienna.

Kirk, J. M. (1996). Corpora and discourse analysis: transcription, annotation, and presentation. In I. Lancashire, C. F. Meyer & C. E. Percy (Eds.), *Synchronic corpus linguistics* (pp. 263-278). Amsterdam: Rodopi.

Kuiper, K. (1991). Sporting formulae in New Zealand English: two models of male solidarity. In J. Cheshire (Ed.), *English around the world: sociolinguistic perspectives* (pp. 200-212). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511611889.014

Kuiper, K., & Flindall, M. (2000). Social rituals, formulaic speech and small talk at the supermarket checkout. In J. Coupland (Ed.), *Small talk* (pp. 183-207). London: Longman.

Kwan, B.S.C. (1996) Introductions in state-of-the-art, argumentative, and teaching tips TESL journal articles: Three possible sub-genres of introductions? *Hong Kong: City University of Hong Kong, Research Monograph,* 12.

Lee, D. Y. W. (2008). Corpora and discourse analysis: new ways of doing old things. In V. K. Bhatia, J. Flowerdew & R. H. Jones (Eds.), *Advances in discourse studies* (pp. 86-99). London & New York: Routledge.

Lewin, B. A., Fine, J., & Young, L. (2001). *Expository discourse: a genre-based approach to social science research texts*. London & New York: Continuum.

Linnakylä, P. (1980). Hi Superman: What is the most functional English for a Finnish five-year-old. *Journal of Pragmatics, 4*, 367-392. doi: 10.1016/0378-2166(80)90031-4

McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Nwogu, K. N. (1990). *Discourse variation in medical texts: Schema, theme and cohesion in professional and journalistic accounts*. Nottingham: University of Nottingham.

Ozturk, I. (2007). The textual organisation of research article introductions in applied linguistics: variability within a single discipline. *English for Specific Purposes, 26*, 25-38. doi: 10.1016/j.esp.2005.12.003

Samraj, B. (2002). Introduction in research articles: Variation across disciplines. *English for Specific Purposes, 21*, 1-18. doi: 10.1016/S0889-4906(00)00023-5

Samraj, B. (2005). An exploration of a genre set: research article abstracts and introductions in two disciplines. *English for Specific Purposes, 24*, 141-156. doi: 10.1016/j.esp.2002.10.001

Schmidt, R. W. (1983). Interaction, acculturation, and the acquisition of communicative competence: a case study of an adult. In N. Wolfson & E. Judd (Eds.), *Sociolinguistics and language acquisition*. (pp. 137-174). Rowley, Mass: Newbury House.

Stubbs, M. (1996). *Text and corpus analysis*. Oxford: Blackwell.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. M. (2004). *Research genres: explorations and applications*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139524827

Swales, J. M., & Feak, C. B. (1994). *Academic writing for graduate students*. Ann Arbor: University of Michigan Press.

Tribble, C., & Scott, M. (2006). *Textual patterns: keyword and corpus analysis in language education*. Amsterdam & Philadelphia: John Benjamins.

Wong Fillmore, L. (1979). Individual differences in second language acquisition. In C. Fillmore, K. Daniel & W. S. Y. Wang (Eds.), *Individual differences in language ability and language behaviour* (pp. 203-228). New York, San Francisco & London: Academic Press.

Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching, 32*, 213-231. doi: 10.1017/S0261444800014154