Becoming a proficient academic writer: shifting lexical preferences in the use of the progressive

Stefanie Wulff[1] and Ute Römer[2]

Abstract

Recent corpus studies have shown that learners of English are aware of systematic associations between verbs and their preferred argument structures to an extent that is similar to that of a native speaker of English (e.g., Gries and Wulff, 2005). Given evidence for similarly systematic associations in native speaker data at the lexis–morphology interface (e.g., Römer, 2005a), the question arises whether, and to what extent, learners of English are also sensitive to lexical dependencies at the level of morphology, and how their verb-aspect associations compare with those of native speakers.

In order to address this question, this study focusses on the potential associations between verbs and progressive aspect in German learners' academic writing. On the basis of the German component of the *International Corpus of Learner English* and the *Cologne–Hanover Advanced Learner Corpus*, learners' significantly preferred verb-aspect pairs are identified using an adaptation of collostructional analysis (Stefanowitsch and Gries, 2003). The results are complemented with corresponding analyses of a subset of the *Michigan Corpus of Upper-level Student Papers* on the one hand and published research articles from the *Hyland Corpus* on the other hand.

The findings indicate that upper-intermediate and advanced German learners of English exhibit clear lexical preferences in the use of progressives. Furthermore, comparative analyses suggest that verb-aspect preferences shift as a function of writers' mastery of text type-specific conventions rather than language proficiency at large.

1. Introduction

This study was triggered by a number of recent findings on the acquisition and use of progressives and other verbal phenomena. To begin with, it is a long-established fact that the acquisition of the progressive constitutes a

[1] Department of Linguistics and Technical Communication, University of North Texas, 1155 Union Circle #305298, Denton, TX 76203–5017, USA.
Correspondence to: Stefanie Wulff, *e-mail:* Stefanie.Wulff@unt.edu
[2] English Language Institute, University of Michigan, 500 E. Washington Street, Ann Arbor, MI 48104–2028, USA.

DOI: 10.3366/E1749503209000276
Corpora Vol. 4 (2): 115–133

challenge for language learners, even at advanced levels, and particularly for learners whose L1 does not have a direct counterpart to the progressive (such as German or Norwegian, for instance; see Johansson and Stavestrand, 1987: 144; and Römer, 2005a: 172). Examples 1 and 2, both taken from in-class essays written by advanced German learners of English, serve to illustrate the misuse of the progressive.[3] As Williams (2002: 18) notes, this language feature also 'constitutes one of the most basic and ubiquitous problems facing language *teachers*' (our emphasis). Therefore, it certainly deserves particular attention in pedagogically orientated corpus research.

(1) We saw the Houses of Parliament and we saw Big Ben. Most people *are thinking* [think] that the tower's name is Big Ben but Big Ben is only the name of the bell.
(2) (Question addressed to a homeless person.) What *are* you *doing* [do you do] every day?

Secondly, recent corpus-linguistic studies suggest that the progressive is not a purely grammatical phenomenon in terms of an empty slot-and-filler model, but that different progressive forms and progressives that express different functions are associated with particular verbs in native speaker data, giving rise to lexical-grammatical patterns (Römer, 2005a, 2005b). These studies on the progressive add to the growing body of converging evidence from corpus linguistics and lexicography (Sinclair, 1991; Stubbs, 1996; and Hunston and Francis, 2000) as well as psycholinguistics and child language acquisition (Goldberg, 2006; and Tomasello, 2003) that testifies to the inseparability of lexis and grammar.

Given the growing evidence for lexical-grammatical patterns in native speaker language, several studies have begun to explore whether, and to what extent, foreign language learners are also sensitive to lexical-grammatical dependencies. Gries and Wulff (2005) addressed this question by conducting two experiments: a syntactic priming and a semantic sorting study. The priming experiment provided evidence that German advanced learners of English can be primed for argument structure constructions like the ditransitive construction or the prepositional dative construction. In the sorting experiment, subjects were given sixteen cards, each with one sentence on them, and were asked to sort these cards into four piles of four cards. The sentences crossed four verbs (*cut*, *get*, *take* and *throw*) and four argument-structure constructions (the transitive, caused-motion, resultative and ditransitive construction). The results showed that subjects significantly preferred a construction-based over a verb-based sorting style. The experimental data were also compared with corpus data from ICE–GB as an L1 corpus and verb-subcategorisation preferences attested in a parsed L1 German corpus (Schulte im Walde, 2003). This revealed that learners'

[3] We would like to thank Sven Naujokat for providing us with these examples from two of his pupils' essays.

verb-specific subcategorisation preferences did not correlate with those attested for German translation equivalents, while they correlated significantly with native English preferences. Accordingly, the results cannot be accounted for by transfer from the L1, but unanimously support the hypothesis that language learners have L2 constructions stored in their interlanguage lexicon (see Gries and Wulff, 2009, for a follow-up study on English complementation constructions).

A fourth point of departure for this study is that previous research has shown that the second language acquisition of the progressive depends on the frequency of the verbs in question as well as the strength of their association with the progressive (as opposed to other tense-aspect morphemes). Verbs that more frequently occur in the progressive and/or are more strongly associated with the progressive in the input are also acquired earlier and produced more frequently in the progressive by learners (Wulff *et al*., 2009). These findings accord nicely with recent studies on the first and second language acquisition of novel constructions: when overall type and token frequencies are held constant, input that is skewed such that one type of example accounts for the preponderance of tokens will result in more accurate generalisation than input that is more representative (Goldberg *et al*., 2004; and Casenhiser and Goldberg, 2005).

By way of synthesising these observations and findings, this study addresses the issue concerning to what extent learners are also sensitive to lexical-grammatical patterns involving the progressive. The genre that we chose to focus on here is that of academic writing since it is a register that is crucial to a large number of language learners. Another motivation was that, to our knowledge at least, the progressive has not been the focus of attention in many studies on academic discourse so far – especially not from a pedagogically minded perspective. This reflects the ongoing trend in the field of academic discourse analysis to concentrate less on core grammatical phenomena and more on metadiscoursal and pragmatic phenomena like hedging, evaluation, argument structure, subjectivity or citation practices (see, for example, Ädel, 2006; Hyland, 1998; Markkanen and Schröder, 1997; Mauranen, 2002; and Römer, 2008).

The central questions we would like to address in the following are how lexical-grammatical patterns vary as a function of language proficiency on the one hand and academic writing expertise on the other. The latter we understand to be inextricably correlated with the thematic foci of the different writing tasks students will be challenged with over their academic career: students begin with argumentative essays before they turn to research papers, lab reports, dissertations, and maybe even publishable papers. Accordingly, developing academic writing expertise can be seen in the shifting vocabulary and grammar choices exhibited in students' writing as they adjust their writing to these different text types and themes.

After a brief overview of the types of data and corpora used in this study, we will discuss the use of progressives by learners, novice native and expert academic writers in terms of their lexical preferences.

GICLE	CHALC	MICUSP_HS	Hyland_HS
upper-intermediate level learner writing	upper-level learner writing	upper-level novice writing	published expert writing
233,849 words	192,364 words	470,766 words	611,209 words

Figure 1: Overview of corpora used in this study

We approach this issue from two complementary perspectives. The first part of the analysis identifies more subtle differences between the four corpora by examining differences in the frequency distribution of selected progressive forms that are attested in all four corpora. The second part seeks to identify sharper differences by establishing which progressives are distinctively associated with either corpus in contrast with the other corpora. In Section 3.3, we highlight some functional differences in the use of the progressive. We conclude with some thoughts on and suggestions for future research.

2. Data: academic writing on different levels of proficiency and writing expertise

We extracted all progressives from four different corpora; Figure 1 provides an overview.

As Figure 1 shows, the data were taken from two foreign language learner corpora that represent two different levels of language proficiency (GICLE and CHALC), and from two corpora that represent novice and expert native-speaker writing respectively (MICUSP and Hyland). The German component of the *International Corpus of Learner English* (GICLE; Granger *et al*., 2002) consists of argumentative essays by (mostly) third-year undergraduate students, totalling about 234,000 words. The second corpus is the *Cologne–Hanover Advanced Learner Corpus* (CHALC; Römer, 2007), which consists of humanities essays and term papers by upper-level students, mainly in their fourth or fifth year, at the universities of Cologne or Hanover, Germany, (i.e., at a level that can be compared to final-year undergraduate and first-year graduate students in the American academic system). The total size of the corpus is about 200,000 words. Thirdly, we considered a sample of 470,000 words from a pre-release version of the *Michigan Corpus of Upper Level Student Papers* (MICUSP).[4] In order to make the corpora as comparable as possible in terms of the essay and

[4] See: http://micusp.elicorpora.info

Corpus	*Types*	*Tokens*	*Words*	*Tokens (pmw)*
GICLE	227	705	233,849	3,014
CHALC	102	245	192,364	1,273
MICUSP_HS	246	697	470,766	1,480
Hyland_HS	294	862	611,209	1,410

Table 1: Distribution of progressives across (sub)corpora

term-paper contents covered, we restricted the MICUSP subset to writing samples from the humanities and social sciences (MICUSP_HS), including linguistics, philosophy, psychology and sociology. MICUSP_HS comprises native and non-native speaker data. While the complete MICUSP contains approximately 20 percent non-native speaker data, only 18 percent of the progressives extracted from MICUSP_HS for this study were produced by non-native speakers. So, overall, the MICUSP data in this study can be taken to represent native speakers for the most part, and advanced foreign language learners of English to a small extent; we subsequently refer to MICUSP_HS as 'novice native speaker writing'. Nevertheless, we will be cautious in our interpretations with regard to what the MICUSP_HS findings tell us about general language proficiency. Finally, we used data from the *Hyland Corpus* (Hyland, 1998), which is a collection of 240 published research articles from eight disciplines. For the purposes of this study, we extracted a sample of around 611,000 words from articles in the fields of linguistics, philosophy and the social sciences (henceforth, Hyland_HS) to match the disciplinary scope of the other three corpora as well as possible. In common with MICUSP_HS, this corpus comprises a number of articles written by non-native speakers of English, so the same caveats with regard to implications for the impact of general language proficiency hold. We can, however, assume that the papers had to pass a native-speaker language check before publication.

From the four selected corpora (containing about 1,515,000 words altogether), we extracted all *–ing* forms, totalling 42,138 hits, which were then checked manually for 'true' hits of progressives. Overall, this yielded 2,509 hits (705 in GICLE, 245 in CHALC, 697 in MICUSP_HS and 862 in Hyland_HS). Table 1 provides an overview of the distribution of progressives across the four (sub)corpora in terms of types, tokens and relative token frequencies per million words (pmw).

If we consider the distribution of progressives across corpora, two things are worth noting. The type and token frequencies in CHALC are rather low (102 and 245, respectively), resulting in a low relative token frequency of 1,273. In contrast, the relative token frequency for GICLE (3,014) is considerably higher than it is for the other three corpora. We will take a closer look at the verbs in the progressive to suggest possible explanations for this observation, below.

Verb	*Pearson residual*	*Verb*	*Pearson residual*
being	–11.492	*going*	1.904
doing	–8.534	*playing*	2.070
using	–6.480	*lying*	2.277
becoming	–6.177	*trying*	2.597
speaking	–5.200	*thinking*	3.077
arguing	–4.459	*working*	5.195
experiencing	–4.179	*drinking*	5.403
taking	–3.584	*feeling*	5.403
leading	–3.208	*losing*	5.403
asking	–2.009	*living*	6.990
telling	–2.009	*coming*	7.265
happening	–1.840	*walking*	9.442
saying	–1.840	*fighting*	11.238
moving	–0.641	*running*	11.238
leaving	–0.432	*getting*	12.000
having	0.306	*standing*	13.030
dealing	0.517	*talking*	14.566
starting	0.570	*sitting*	26.852
looking	1.553		

Table 2: Pearson residuals GICLE compared to Hyland_HS in ascending order

3. Results: lexical and functional preferences in novice and expert use of progressives

3.1 Lexical preferences I: progressive frequencies

A cursory glance at the raw data reveals that not all verbs occur across all four corpora – as a matter of fact, a number of them (369 types in total) occur only in one corpus. In this section, we begin with an examination of the distribution of those verbs that occur in all four corpora; we devote Section 3.2 to an analysis of the total set of verbs.

In what follows, the basis of comparison are the frequencies observed in Hyland_HS, since this corpus represents the target distribution in the sense that it is a representative sample of expert academic writing that the other corpora are assumed to be working towards. For the thirty-seven verbs that occur in the progressive in all four corpora, chi-square tests were calculated to identify differences in the frequency distributions of those verbs in Hyland_HS compared to the other three corpora, respectively. Let us look at each comparison in turn, beginning with the upper-intermediate level non-native writers in comparison with the expert writers (GICLE). A chi-square test on the absolute frequencies of the thirty-seven verbs in question turns out to be very highly significant ($\chi^2 = 2225.33$; df = 36; $p < .001$***). Table 2 displays the Pearson residuals of the GICLE frequencies in contrast to

Verb	*Pearson residual*	*Verb*	*Pearson residual*
being	–7.554	*trying*	1.133
doing	–7.009	*drinking*	1.358
becoming	–4.358	*fighting*	1.358
taking	–3.843	*leaving*	1.358
happening	–3.438	*losing*	1.358
using	–2.986	*running*	1.358
arguing	–2.602	*living*	1.510
saying	–2.337	*moving*	1.510
having	–1.860	*standing*	2.895
lying	–1.807	*coming*	3.483
working	–1.807	*playing*	3.670
leading	–1.489	*walking*	4.375
speaking	–1.236	*starting*	5.850
thinking	–1.015	*sitting*	7.299
experiencing	–1.009	*talking*	9.260
looking	0.046	*going*	9.932
telling	0.123	*asking*	10.441
dealing	0.228	*feeling*	14.027
getting	0.228		

Table 3: Pearson residuals CHALC compared to Hyland_HS in ascending order

Hyland_HS. A high negative residual means that the verb in question occurs less frequently than expected in GICLE than in Hyland_HS; high positive residuals mean that they occur more frequently.

Contrasting Table 2 with Tables 3 and 4, which show the Pearson residuals for CHALC and MICUSP_HS, we see first of all that the verb frequencies in GICLE deviate most strongly from Hyland_HS: especially for the verbs occurring more frequently in GICLE, the Pearson residuals are the highest overall (r_{Pearson} *sitting* $= 26.852$; *being* yields the lowest residual: $r_{\text{Pearson}} = -11.492$). Secondly, these verbs form a quite coherent semantic group: *sitting*, *standing*, *running*, *fighting*, *walking* and *coming* are all among the verbs with the highest residuals, and they are all physical action verbs. At the other end of the spectrum, we find the two bleached verbs *being* and *doing*. Another group of verbs that are comparatively underrepresented in the GICLE data are communication verbs such as *speaking*, *arguing*, *asking*, *telling* and *saying* (with one important exception: *talking* yields the second highest positive residual in GICLE compared to Hyland_HS; $r_{\text{Pearson}} = 14.566$). So, overall, we see a clear bias in the upper-intermediate learner data towards concrete, physical action verbs, while general and communication verbs are dispreferred. An initial interpretation could be to ascribe these preferences to the typical thematic content of argumentative essays as covered in GICLE. However, let us first turn to the other corpora.

Table 3 displays the Pearson residuals obtained from the advanced level learners (CHALC) compared with the expert writers.

Table 3 reveals a slightly different picture from that found in the GICLE data. First of all, we can note that in terms of absolute values, the Pearson residuals for the advanced learners are lower than for the upper-intermediate learners (r_{Pearson} *being* $= -7.554$; r_{Pearson} *feeling* $= 14.027$). In other words, the frequency distribution of the verbs differs less starkly from that in Hyland_HS overall. It is also, however, very highly significant ($\chi^2 = 834.67$; df $= 36$; $p < .001$***). With regard to the verbs that are comparatively frequent, we also find a number of physical action verbs (such as *sitting* and *walking*), but these verbs do not have such high residuals in the CHALC data as they have in the GICLE data. Also, physical action verbs are not the most frequent verbs here, but, instead, we find the perception verb *feeling*, followed by the communication verb *asking*. It appears that, overall, communication verbs are used slightly more frequently in CHALC than in GICLE when being compared with the target corpus Hyland_HS – note, for instance, the slightly positive residual value for *telling* (0.123), which indicates that the frequency of this verb is highly similar to that in Hyland_HS; *telling* yielded a decidedly more negative residual in the GICLE data (–2.009). However, verbs such as *saying* and *speaking* have negative residuals in the CHALC data, too (–2.337 and –1.236, respectively). Overall, this testifies not to a dramatic difference between the upper-intermediate and the advanced learners, but to a rather subtle one. Another similarity between the two learner data sets concerns the under-representation of the general purpose verbs *being* and *doing*, which again obtain the lowest residuals (-7.554 and -7.009, respectively).

But how different are the frequency distributions of novice (predominantly) native speaker writers compared to the expert writers? If we consider Table 4, we see that the overall distribution is very highly significant again, but even less so than the other two ($\chi^2 = 355.2415$; df $= 36$; $p < .001$***). That the MICUSP_HS distribution more closely approximates that in Hyland_HS is also reflected in the even smaller range of values that the Pearson residuals take on (r_{Pearson} *doing* $= -4.538$; r_{Pearson} *working* $= 9.053$). With regard to the verbs occurring more often than expected, a picture emerges that is quite reminiscent of that observed in the GICLE learner data: the most frequent verbs (again, in specific contrast to Hyland_HS, we need to bear in mind) are concrete verbs denoting physical actions. *Working* yields the highest residual of 9.053; other examples include *moving* (3.799), *running* (2.775), *drinking* (2.775), *walking* (1.374) and *fighting* (1.374). As in the learner data, the one communication verb that is clearly over-represented is *talking* (8.724). *Telling*, which scored slightly higher in the advanced foreign language learner data compared with the upper-intermediate foreign language learners data, yields an even higher value in MICUSP_HS (2.197). However, we see that the other communication verbs, such as *arguing*, *saying* or *asking*, are used less frequently than expected by novice native speaker writers, just as is the case with our two learner groups.

Verb	*Pearson residual*	*Verb*	*Pearson residual*
doing	–4.538	*playing*	0.325
becoming	–3.902	*using*	0.450
looking	–3.010	*trying*	0.818
dealing	–2.751	*going*	1.138
lying	–2.155	*coming*	1.287
speaking	–2.078	*fighting*	1.374
leading	–1.922	*leaving*	1.374
thinking	–1.863	*walking*	1.374
having	–1.442	*getting*	1.730
sitting	–1.371	*telling*	2.197
happening	–1.096	*experiencing*	2.760
taking	–0.917	*drinking*	2.775
standing	–0.605	*losing*	2.775
starting	–0.605	*running*	2.775
asking	–0.424	*moving*	3.799
being	–0.387	*feeling*	7.679
saying	–0.073	*talking*	8.724
arguing	–0.062	*working*	9.053
living	–0.053		

Table 4: Pearson residuals MICUSP_HS compared to Hyland_HS in ascending order

Overall, then, with regard to the most noticeable patterns to be observed in verbs that occur in all four corpora, the results so far suggest that their frequency distribution is not just a matter of thematic focus or general language proficiency, but indicate that another strong predictor of how often a specific verb will be employed is academic writing expertise in particular.

However, the analysis so far (deliberately) disregarded a large number of verb types that do not occur in all four corpora, and, in many ways, it can be argued that it is these verbs that may shed even more light on the differences between the four corpora. What if, for example, there are verbs that occur only in MICUSP_HS and Hyland_HS, but never or only rarely in the foreign language learner data sets? Such a finding would put into perspective the impression we have so far of rather insignificant differences as far as native speaker status is concerned. We turn, therefore, to a more comprehensive analysis of all verbs attested at least once in at least one of the corpora in the following section.

3.2 Lexical preferences II: distinctive progressives

While the analysis of verbs shared among the four corpora has already hinted at some interesting differences, this section adopts a slightly more quantitative perspective and asks, ‘When does a verb occur sufficiently

frequently in a corpus to license the conclusion that it is *distinctly associated* with that corpus?'

In order to address this issue, we subjected the four selected datasets to a so-called Distinctive Collexeme Analysis (DCA). DCA is a member of the family of collostructional analyses developed by Gries and Stefanowitsch (see Gries and Stefanowitsch, 2004). The most basic application of that family of methods is collexeme analysis, an extension of the concept of significant collocates to co-occurrences not just of two words, but of words and other linguistic elements, most notably syntactic patterns or constructions. Lexemes that are significantly associated with a construction are referred to as collexemes of that construction. The association is quantified by means of the log to the base of ten of the *p*-value of the Fisher Yates exact test (see Stefanowitsch and Gries, 2003: 217–8, for justification).

As an extension of collexeme analysis, DCA specifically compares two or more closely related, or even largely synonymous, constructions. DCA has so far mostly been applied to look into the association between words and constructional variants, such as the dative alternation or particle placement. For the purpose of this study, we use it to identify the progressives that are distinctive (i.e., typical) for the four corpora that we consider here. All computations were done with Stefan Gries's (2004) *R*-script coll.analysis 3.[5] The script uses an exact binomial test to quantify the strength of the association between progressives and the corpus in which they occur. More precisely, it provides a *p*-value for each progressive and log-transforms it such that highly positive and highly negative values indicate a large degree of attraction and repulsion respectively, while 0 indicates random co-occurrence. An (absolute) p_{log} value that is equal to or higher than 1.3 corresponds to a probability of error of 5 percent or less.

Table 5 displays the progressives that are most significantly distinctive in any of the four corpora (in descending order of their strength of association).

Let us take a closer look at the significantly associated progressives for each corpus in turn. Beginning with the upper-intermediate learner corpus, GICLE, we can first of all observe a strong preference for physical action verbs (*riding*, *jumping*, *driving* and *running*), many of which also imply repetitive motion. In CHALC, we also find some examples of physical motion verbs, although *dancing* is the only one that is above the significance level of 1.3. Looking at the most distinctive progressives in MICUSP_HS, we find no such bias towards physical action verbs (only *traveling* may be taken to at least imply physical motion). This predominance of motion verbs again points to the generic focus of GICLE on argumentative essays in which students often draw on their personal experience and tell anecdotes from their personal lives. However, there are clear indications that non-nativeness or lack of expertise has an effect as well: we can observe a strong tendency in the

[5] The script is available from the author's website, at: http://www.linguistics.ucsb.edu/faculty/stgries/research/overview-research.html (Accessed 22 May 2008).

GICLE	$p_{\log}$	*CHALC*	$p_{\log}$	*MICUSP_HS*	$p_{\log}$	*Hyland_HS*	$p_{\log}$
sitting	7.29	*going*	6.74	*measuring*	3.89	*doing*	3.63
riding	3.31	*asking*	4.05	*violating*	3.34	*keeping*	3.14
getting	3.20	*presenting*	2.46	*pursuing*	2.78	*operating*	2.78
watching	3.07	*corresponding*	2.02	*expressing*	2.65	*acting*	2.76
wearing	3.01	*heading*	2.02	*conceiving*	2.23	*being*	2.54
jumping	2.76	*quoting*	2.02	*working*	2.20	*becoming*	2.52
missing	2.76	*feeling*	1.81	*learning*	2.15	*claiming*	2.32
standing	2.67	*boiling*	1.57	*facing*	1.69	*starving*	2.32
waiting	2.31	*starting*	1.45	*affecting*	1.67	*considering*	2.31
chatting	2.21	*dancing*	1.30	*behaving*	1.67	*beginning*	1.97
ringing	2.21	*dreaming*	1.30	*mapping*	1.67	*reading*	1.97
fighting	1.83			*signaling*	1.67	*alluding*	1.86
living	1.72			*testing*	1.67	*proposing*	1.69
talking	1.68			*traveling*	1.67	*discovering*	1.39
cooking	1.65			*establishing*	1.64	*embarking*	1.39
cycling	1.65			*expecting*	1.64	*forthcoming*	1.39
decreasing	1.65			*perceiving*	1.64	*introducing*	1.39
lacking	1.65			*seeing*	1.64	*supposing*	1.39
shining	1.65			*creating*	1.64	*developing*	1.31
trembling	1.65			*experiencing*	1.42	*pulling*	1.30
driving	1.62			*being*	1.33	*reaching*	1.30
growing	1.62					*seeking*	1.30
running	1.57					*focusing*	1.30
thinking	1.40					*suggesting*	1.30

Table 5: Top distinctive progressives for the four (sub)corpora in descending order of distinctiveness

GICLE data towards the use of inherently stative verbs in progressive form (*missing*, *lacking* and *decreasing*), which is an obvious overgeneralisation of the 'ongoing event'-interpretation of the progressive. Examples 3 to 5 testify to the learners' attempts to make use of this meaning potential of the progressive.

(3) We *will* surely *be missing* something when we withdraw from the daily confrontation with other human beings. (GICLE)
(4) Some skinny people *are lacking* strength and health. (GICLE)
(5) Hundreds of Spanish actors play their story of the ship named Olympic Idea on the ocean of life, which *is endangering* it. (GICLE)

Turning towards the CHALC data (i.e., advanced German learner production), we see that, contrary to GICLE, the most distinctive progressives relate to an academic context (consider *presenting*, *corresponding* and *quoting*); this result again reflects the types of text included in the corpus (essays and term papers rather than argumentative essays). However, while these verbs are much closer to what we would expect to find in an academic piece of writing, the fact that they are used in the progressive seems to require further explanation. A look at the CHALC concordances of these forms reveals that the advanced learners of English also overgeneralise the progressive, if only with more target-like verbs, as in Examples 6 to 8.

(6) Sometimes it seems as if the authors *are presenting* their own subtheories under the cover of functional grammar. (CHALC)
(7) You use 'say' when you *are quoting* directly the words someone has spoken. (CHALC)
(8) The second *are corresponding* to both, yes and no, accompanying different degrees of frequency, such as sometimes yes, sometimes no. (CHALC)

Let us now turn to the novice native speaker writers in MICUSP_HS. While these are upper-level students, we do not find many reporting verbs that are typical of academic writing; instead, whereas the most distinctive verbs clearly relate to an academic context of scientific research, they denote concrete actions involved in research, such as *measuring*, *signaling* and *testing*. This establishes an interesting contrast with the CHALC data, which capture lower general language proficiency, but arguably a higher academic writing proficiency. In CHALC, we find a number of reporting verbs like *presenting* and *quoting*, none of which are distinctive in MICUSP_HS writing. That is, the novice native speaker writers' focus seems to be less on the abstract examination of a topic and more on the actual processes, experiments, and tests involved in carrying out research. Examples 9 and 10 illustrate this use.

(9) The validity of a test is an index of the extent to which the test *is measuring* what it is hoped to measure. (MICUSP_HS)

(10) If what the Rorschach *is testing* can be specified, it will be able to render useful information other tests cannot offer… (MICUSP_HS)

The DCA also confirms our initial observation that *being* is distinctive for MICUSP_HS and Hyland_HS, which runs counter to the expectation that this high-frequency verb may be used much more by learners as a fallback for the lack of more specific alternatives.[6] Looking into the Hyland data, we see that quite the contrary is true. The tendency for high-frequency, semantically bleached verbs is most prominent in published academic writing (when being contrasted with the other three corpora): *doing*, *keeping* and *becoming* all rank among the top distinctive progressives in this corpus.

In addition, there are various verbs in Hyland_HS that obviously perform a metadiscoursal function, meaning that they are used to introduce, refer to, and discuss either the research described in the article itself, or that carried out by other researchers and presented in other publications. Verb forms like *claiming*, *considering*, *alluding*, *proposing*, *discovering*, *embarking* and *introducing* are examples. In other words, when we look at the differences between our novice writers in MICUSP_HS and the expert writers in Hyland_HS, concrete procedure verbs have given way to more abstract verbs like *proposing*, *developing* and *focusing*. Examples 11 to 13 show such verbs in their context.

(11) Those who *are developing* expertise in writing already command a first-roal language … (Hyland_HS)
(12) In any case I *am* not *proposing* this alternative, but merely examining its implications. (Hyland_HS)
(13) … gives the impression he *is claiming* that in late modernity individuals' knowledgeability has become increasingly self-conscious and discursive. (Hyland_HS)

To summarise, while the verbs distinctively associated with the novice native speaker writers are, arguably, more specifically related to the academic domain than the few reporting verbs we find for CHALC (which is partly a problem of corpus size – larger data samples may well reveal a different picture), they resemble more closely the upper-intermediate learners in their focus on concrete processes. In other words, the data suggest a parallel shift in both non-native and native speaker writing from more concrete, physical action verbs to abstract, metadiscoursal verbs.

---

[6] As one reviewer pointed out, the underuse of bleached verbs by learners may also well be teaching-induced since learners are often told not to use stative verbs in the progressive. It is quite interesting to see, then, that while the learners apparently adhere to this rule for the high frequency verbs, they still 'violate' it with less frequent stative verbs, as we saw earlier in Examples 3 to 8.

Time reference	*GICLE*	*CHALC*	*MICUSP_HS*	*Hyland_HS*	*Total*
Future	14 (9)	1 (3)	12 (9)	6 (11)	33
Past	**219 (169)**	37 (59)	**92 (167)**	253 (206)	601
Present	472 (527)	207 (183)	593 (521)	603 (644)	1,875
Total	705	245	697	862	2,509

Table 6: Distribution of time reference parameters across the four (sub)corpora

3.3 Preferences in terms of time reference and modality

In a next analytic step we submitted our datasets to a functionally orientated classification and determined for each of the 2,509 progressive forms extracted from our four corpora whether they expressed past, present or future time reference. We then compared the findings based on the four (sub)corpora. Table 6 provides an overview of the results showing the observed frequencies, expected frequencies in parentheses, and (near-) significant contributions to chi-square highlighted in boldface.

As we can see in Table 6, the distribution of time reference types across corpora turned out to be very highly significant ($\chi^2 = 95.810$; df = 6, $p < .001$***). There is a tendency towards past tense reference in both GICLE and MICUSP_HS (the latter being much more pronounced than the former, which marginally misses the threshold for significance, but constitutes the second-highest contribution by far, which is why it is considered here). Looking at the concordance lines more closely, we find that both are reflections of the writing tasks at hand: in the GICLE data, past tense is mainly employed in the personal narrative context, as shown in Examples 14 to 16.

(14) They *had been drinking* a lot and shortly after midnight they saw that they were running out of wine. (GICLE)
(15) I *had been watching* the street intently all morning waiting for the postman … (GICLE)
(16) I *was watching* a crime story that evening … (GICLE)

In the MICUSP_HS data, we find that the novice writers fall back on a similarly narrative style to make reference to previous research, or to lay out how they went about their own research project. Examples 17 to 19 demonstrate this.[7]

[7] The sceptical reader may ask whether or not this similarity between GICLE and MICUSP_HS can be attributed to the non-native speakers in MICUSP_HS. We ran this analysis for the native speakers and non-native speakers in MICUSP_HS separately and can confirm that the results are virtually identical.

(17) One group of researchers *were examining* the question from an information-processing perspective, … (MICUSP_HS)
(18) Sula *was* simply *trying out* in her life what she had witnessed through direct modeling. (MICUSP_HS)
(19) In studies on the Canadians, lexical replacements *were overtaking* phonological and pronunciation variants by almost a 2 : 1 ratio. (MICUSP_HS)

On the contrary, the Hyland_HS data are less distinctly associated with past time reference than GICLE and MICUSP_HS. We may speculate that this is due to the fact that in a published research article, as opposed to a student research paper, less space will be reserved for literature review and procedural explanations, and more space will be devoted to the presentation of results and discussion, which may be the ultimate reason for this tense bias. Also, there are various examples in Hyland_HS in which reference to previous work by other researchers is made in the present tense, thereby foregrounding an idea or concept as a permanent thing (as opposed to the research process leading up to that idea or concept); Examples 20 to 22 show this use of the present tense.

(20) Wittgenstein (1968, 102) *is* clearly *alluding* to Kant and transcendental philosophy (and to his own earlier work) … (Hyland_HS)
(21) … a position similar to the one Craib *is proposing*. (Hyland_HS)
(22) Rescher and Brandom *are modeling* inference for logic and mathematics, and possibly for metaphysics. (Hyland_HS)

So, the parallels between the GICLE and MICUSP_HS data with regard to preferred time reference of progressive forms, paired with a quite different picture in Hyland_HS, suggest that the primary variable at work is the thematic focus of the writing task at hand.

Finally, let us look briefly at another possible function of the progressive as a marker of modality, namely to express a modal-like meaning, (i.e., a meaning that is 'associated with obligation, necessity, possibility, and other aspects of modal meaning', see Hunston, 2008: 272). Examples 11 to 13 exemplify this use in expert academic writing: in these examples, specifically, the progressive arguably functions as a hedging device; in Example 12, it adds emphasis to the negation. How prominent is this function in the other data sets? While a detailed analysis of all 2,509 attestations is beyond the scope of this paper, we can reach some preliminary conclusions by considering all occurrences of progressives as part of a modal frame; that is, a pattern that also features a modal verb or a lexical verb performing a modal function. Table 7 provides an overview of all instances of such modal frames attested in the data.

While the frequencies are too small to license far-reaching conclusions (and also bearing in mind that modality need not be expressed

GICLE	*seem to be* (6), *might be* (3), *should be* (2), *would be* (2), *may be* (1)
CHALC	*would be* (3), *could be* (2), *seem to be* (2), *appear to be* (1), *should be* (1)
MICUSP_HS	*may be* (20), *should be* (8), *would be* (7), *could be* (6), *seem to be* (6), *likely to be* (4), *might be* (3), *appear to be* (2), *claim to be* (1), *ought to be* (1), *rumored to be* (1), *said to be* (1)
Hyland_HS	*would be* (12), *may be* (8), *might be* (5), *seem to be* (5), *appear to be* (4), *seen to be* (2), *claim to be* (1), *could be* (1), *interpret to be* (1), *likely to be* (1), *ought to be* (1), *perceive to be* (1), *propose to be* (1), *take to be* (1), *understand to be* (1)

Table 7: Modal frames followed by progressive verb forms across the four (sub)corpora

exclusively in modal frames like the ones in Table 7), in terms of the variety of patterns, it is quite obvious that both foreign language learners and novice native speaker writers employ only a fraction of the modal frames that the expert writers in Hyland_HS have at their disposal. Moreover, both foreign language learners and novice native speaker writers rely on high frequency modal verbs for the most part. That *ought to be* occurs in MICUSP_HS, but neither GICLE nor CHALC, is due to varietal differences (foreign language instruction in Germany is usually much more strongly orientated towards British English). What we can conclude from this snippet is that all novice writers (native and non-native speakers) may benefit from writing instruction that brings the modal use of the progressive into focus, particularly as part of modal frames like the above.

4. Conclusion and outlook

Several interesting findings emerge from the corpus analyses reported on in this paper. First, we found that there are systematic verb-progressive associations in academic writing, and these preferences shift systematically as the writing tasks move from verbs denoting physical action to metacommunication verbs. This is accompanied by a corresponding functional shift of the progressive from a 'continuous single event'-reading (i.e., the meaning that is put forward as the core progressive meaning in EFL teaching materials, see Römer, 2005a) to a more modal meaning. This functional shift is, in turn, accompanied by a grammatical shift from strong biases towards (narrative and personal) past tense to (fact-oriented and objective) present tense usage. Together, these findings suggest that both the genre-specific verb preferences for progressives and their functions in academic writing need to be acquired by novice writers regardless of their native speaker status. That is, the acquisition of these genre-specific lexical-grammatical patterns is less a matter of language proficiency in general,

but, rather, more one of writing proficiency in particular. In other words, it is not so much *nativeness* that affects lexical-grammatical choices in academic writing than *expertise* in a certain genre (see Römer, 2009, for further supportive evidence).

At the same time, it was interesting to see that the progressives most strongly associated with published writing are highly frequent, semantically bleached verbs. These tend to be avoided by novice writers, perhaps in an attempt to strike the right note by using verbs that are allegedly more target-like. It is only by examining the *conditional* use of verbs, as being tied to a particular grammatical structure and as being situated in a particular genre, that tendencies like these become transparent. Consequently, this analysis may serve as one example of how corpus-linguistic methods like collostructional analysis can help us gain a more precise understanding of genre-specific lexical-grammatical dependencies.

In conclusion, this study adds to the growing body of evidence that even core grammatical phenomena like the progressive are highly genre-dependent (see, for example, Hyland, 1998, 2000; and Swales, 1990, 2004), and that meaning and function associations permeate all layers of language, here including the choice of tense and aspect, lexical items, and larger frames with conventionalised functions. This calls for further research on the impact of genre on lexical-grammatical choices, including the implications of the results for EAP teaching. Future research could go beyond the scope of the analysis here in various ways. For one, future studies should include learner data from L1 backgrounds other than German; cover a wider range of academic disciplines; and consider all verb forms of the verb lemmata involved. One problem we see in this context is the availability (or unavailability, as the case may be) of suitable and sufficiently large corpora that capture learner and novice academic writing at different levels of proficiency. We hope for larger learner corpora that control for proficiency levels to complement the *International Corpus of Learner English*. Similarly, we think it would be desirable to have a larger corpus of published academic writing with wide disciplinary and genre coverage that could function as a more representative reference corpus.

The current limitations aside, we hope to have illustrated the potential of a corpus-linguistic approach to shed light on the development of genre-dependent language proficiency. We also hope that our paper will stimulate more related, applied corpus-linguistic research that may, in the long run, help novice writers to improve their academic writing proficiency and thus become better members of the academic community of practice.

References

Ädel, A. 2006. Metadiscourse in L1 and L2 English. Amsterdam/ Philadelphia: John Benjamins.

Casenhiser, D. and A.E. Goldberg. 2005. 'Fast mapping of a phrasal form and meaning', Developmental Science 8 (6), pp. 500–8.

Goldberg, A.E. 2006. Constructions at Work: The Nature of Generalization in Language. Oxford: Oxford University Press.

Goldberg, A.E., D. Casenhiser and N. Sethuraman. 2004. 'Learning argument structure generalizations', Cognitive Linguistics 14 (3), pp. 289–316.

Granger, S., E. Dagneaux and F. Meunier (eds). 2002. International Corpus of Learner English. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.

Gries, St. Th. 2004. Coll.analysis 3. A Program for R for Windows.

Gries, St. Th. and A. Stefanowitsch. 2004. 'Extending collostructional analysis: a corpus-based perspective on "alternations" ', International Journal of Corpus Linguistics 9 (1), pp. 97–129.

Gries, St. Th. and S. Wulff. 2005. 'Do foreign language learners also have constructions? Evidence from priming, sorting and corpora', Annual Review of Cognitive Linguistics 3, pp. 182–200.

Gries, St. Th. and S. Wulff. 2009. 'Psycholinguistic and corpus-linguistic evidence for L2 constructions', Annual Review of Cognitive Linguistics 7, pp. 164–87.

Hunston, S. and G. Francis. 2000. Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English. Amsterdam: John Benjamins.

Hunston, S. 2008. 'Starting with the small words: patterns, lexis and semantic sequences' in U. Römer and R. Schulze (eds) Patterns, Meanings and Specialized Discourses. Special issue of the International Journal of Corpus Linguistics 13 (3), pp. 271–95.

Hyland, K. 1998. Hedging in Scientific Research Articles. Amsterdam: John Benjamins.

Hyland, K. 2000. Disciplinary Discourses: Social Interactions in Academic Writing. London: Longman.

Johansson, S. and H. Stavestrand. 1987. 'Problems in learning – and teaching – the progressive form' in I. Lindblad and M. Ljung (eds) Proceedings of the Third Nordic Conference for English Studies, Hässelby, 25–27 September 1986, Volume 1, pp. 139–48. Stockholm: Almquist and Wiksell.

Markkanen, R. and H. Schröder (eds). 1997. Hedging and Discourse: Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts. Berlin: Walter de Gruyter.

Mauranen, A. 2002. "'A good question." Expressing evaluation in academic speech' in G. Cortese and P. Riley (eds) Domain-specific English: Textual Practices across Communities and Classrooms, pp. 115–40. Berne: Peter Lang.

Römer, U. 2005a. Progressives, Patterns, Pedagogy. A Corpus-driven Approach to English Progressive Forms, Functions, Contexts and Didactics. Amsterdam: John Benjamins.

Römer, U. 2005b. 'Shifting foci in language description and instruction: towards a lexical grammar of progressives', Arbeiten aus Anglistik und Amerikanistik 30 (1), pp. 145–60.

Römer, U. 2007. 'Learner language and the norms in native corpora and EFL teaching materials: a case study of English conditionals' in S. Volk-Birke and J. Lippert (eds) Anglistentag 2006 Halle. Proceedings, pp. 355–63. Trier: Wissenschaftlicher Verlag Trier.

Römer, U. 2008. 'Identification impossible? A corpus approach to realisations of evaluative meaning in academic writing', Functions of Language 15 (1), pp. 115–30.

Römer, U. 2009. 'English in academia: does nativeness matter?', Anglistik: International Journal of English Studies 20 (2).

Schulte im Walde, S. 2003. Experiments on the Automatic Induction of German Semantic Verb Classes. Doctoral dissertation, University of Stuttgart. (Arbeitspapiere des IMS 9.2, Universität Stuttgart.)

Sinclair, J. 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Stefanowitsch, A. and St. Th. Gries. 2003. 'Collostructions: investigating the interaction between words and constructions', International Journal of Corpus Linguistics 8 (2), pp. 209–43.

Stubbs, M. 1996. Text and Corpus Analysis: Computer-assisted Studies of Language and Culture. Oxford: Blackwell.

Swales, J.M. 1990. Genre Analysis: English in Academic and Research Settings. Cambridge: Cambridge University Press.

Swales, J.M. 2004. Research Genres: Exploration and Applications. Cambridge: Cambridge University Press.

Tomasello, M. 2003. Constructing a Language. A Usage-based Theory of Language Acquisition. Cambridge, Massachusetts: Harvard University Press.

Williams, C. 2002. Non-progressive and Progressive Aspect in English. Fasano: Schena.

Wulff, S., N.C. Ellis, U. Römer, K. Bardovi-Harlig and C.J. LeBlanc. 2009. 'The acquisition of tense-aspect: converging evidence from corpora, cognition, and learner constructions', Modern Language Journal 93 (3), pp. 354–69.