UTE RÖMER, Ann Arbor

English in Academia: Does Nativeness Matter?

1. Introduction: English as an Academic Language

When we think about 'non-native Englishes', academic English may not be the first thing that comes to mind. Other types of Englishes which tend to exhibit non-standard features and deviate more clearly from the native norm (or from a set of different native norms), e.g. learner Englishes and ESL varieties are perhaps more prototypical candidates. In academia, we are dealing with high-proficiency English speech and writing in most branches. We do not normally find non-standard features like missing articles or third-person-singular '-s' in academic English as represented in research articles, academic lectures, doctoral dissertations or book reviews – and yet, it is a fact that a large and growing number of these (and other) types of text in academic contexts are produced by non-native speakers of English. As Swales (2004, 43) puts it, "in research and scholarship", we seem to be "approaching a situation in which English is becoming a genuine *lingua franca*." In a lecture on "Writing in the academy", Hyland (2006), too, refers to the dramatically increasing numbers of academics whose L1 (first language) is not English but who publish in English (see also Bruce 2008 and Flowerdew 2007).

So, if this is true, if the research world is becoming more and more Anglicized and large numbers of non-native speakers (or "non-Anglophones", in Swales' 2004, 46 terms) produce academic English alongside their native-speaker colleagues, what status does nativeness have in this context? In other words, does nativeness matter when we are dealing with English in academia or are there other, perhaps more important aspects to consider that influence our performance in academic English settings? The present article will address this question by examining what the native/non-native-distinction means in the context of English academic writing. It will investigate how different the academic writing of native speakers and non-native speakers of English is and, based on comparisons of apprentice and expert performance data (in Bazerman's 1994, 131 terms), discuss whether nativeness has an effect on academic writing proficiency if other potentially influential factors like genre, discipline, and duration of university education are controlled.

The focus of the analyses will be on frequent phraseological items, e.g. word combinations such as *on the one hand* or *in the case of* that are typical of academic writing, in comparable sets of successful apprentice academic writing (henceforth AAW) by native speakers and non-native speakers of English in the disciplines of Linguistics and English (language and literature). A collection of published expert academic writing (research articles from Linguistics journals) will function as a reference corpus and will be regarded as a kind of target norm for our apprentice writers.

***Anglistik: International Journal of English Studies* 20.2 (September 2009): 89-100.**

2. Case Study: Exploring the Phraseological Profile of Apprentice Academic Writing (AAW)

2.1 Data and Method

For the case study on phraseological items that constitutes the centre of this paper, I collected data from two corpora of apprentice academic writing (AAW, by native and non-native speakers of English) and from a comparable corpus of expert academic writing (EAW, mainly by native speakers of English; see below). Following Scott and Tribble (2006, 133), apprentice writing samples are understood to be "unpublished pieces of writing that have been written in educational or training settings", whereas expert writing samples are pieces of writing that have been published. The three corpora I used were:

(i) CHALC, the Cologne-Hanover Advanced Learner Corpus (Römer 2007), a collection of currently 45 essays and term papers written by upper-level university students (mainly final year undergraduates and first year graduates, L1 = German) at the universities of Cologne and Hanover, Germany, in English linguistics and English Literary Studies; size: ~ 200,000 words;

(ii) MICUSP_EL, a subset of 91 English (language and literature) and Linguistics papers (term papers, essays, literature reviews) of the Michigan Corpus of Upper-level Student Papers (under compilation at the University of Michigan English Language Institute, see http://elicorpora.info); the subset only covers writing samples by senior (i.e. final year) undergraduate students whose first language is English; size: ~ 200,000 words;

(iii) Hyland_Ling, the Linguistics subsection (30 published research articles from the field of Linguistics) of the Hyland Corpus (Hyland 1998); size: ~ 210,000 words.

CHALC and MICUSP_EL are very similar in terms of disciplinary, student level, and text type coverage so that the only major difference between the two corpora (and the variable under scrutiny in this paper) is the student writers's native-speaker status. What distinguishes Hyland_Ling from CHALC and MICUSP_EL are the level of academic writing expertise of the authors who contributed to the corpora and the number of years they spent in academia. All Hyland_Ling authors are, in Swales' (2004, 56) terms, "senior" scholars, have had multiple-year university training, and managed to get published in peer-reviewed journals. While not all Hyland_Ling authors are native speakers of English, it can, however, be assumed that articles by non-native speakers have been checked and corrected by a native speaker. This is at least the policy of the journals from which the Hyland_Ling articles have been taken. Even though 'academic discipline' is a controlled variable in our data collection, with all three corpora consisting of texts from Linguistics and/or English (language and literature), the topics dealt with in the papers included in CHALC, MICUSP_EL and Hyland_Ling are quite varied and range from accounts of Functional Grammar and generative syntax to discussions of language acquisition phenomena and interpretations of Shakespeare's plays or Milton's poetry. I will take this into account in the discussion of the findings.

In the exploration of the phraseological profile of native and non-native speaker AAW and EAW, a new-generation corpus tool that I would label a 'phraseological search engine' was used: *kfNgram* (Fletcher 2002-2007). *KfNgram* is a program that extracts lists of n-grams of different lengths (i.e. combinations of n words) from a corpus. For the present study I extracted frequency lists of 3-grams (such as *the use of*), 4-grams (such as *at the same time*) and 5-grams (such as *at the end of the*) from CHALC, MICUSP_EL and Hyland_Ling. The analyses below, however, are only based on 3- and 4-grams, since the 5-gram lists were comparatively short and mainly consisted of topic-related items, such as *the specifier position of the* or *the social psychology of language*.

In addition to producing lists of repeated word combinations, *kfNgram* also identifies patterns in the extracted n-gram lists and groups n-grams that differ by only one word in the same position together, e.g. *at the end of*, *at the beginning of* and *at the risk of.* Such groups of n-grams are called phrase-frames (short 'p-frames') and contain a wildcard character (*) that replaces any one word. The p-frame *at the * of* thus summarizes the 4-grams *at the end of*, *at the beginning of* and *at the risk of*, all variants of *at the * of.* Together with the types and token numbers of the p-frames, *kfNgram* lists how many variants are found for each of the p-frames. A p-frame analysis hence provides insights into pattern variability and helps us see to what extent John Sinclair's Idiom Principle (Sinclair 1987, 1991, 1996) is at work, i.e. how fixed language units are or how much they allow for variation.

The identification of n-grams that are specific to a particular collection of texts (e.g. CHALC as compared to MICUSP_EL, or MICUSP_EL as compared to Hyland_Ling) was facilitated by the program *WordSmith Tools* (version 4.0, Scott 2004). I carried out an n-gram frequency comparison using the 'compare 2 wordlists' function in the *WordList* tool. The lists I compared, however, did not consist of single words (i.e. 1-grams) but of sequences of 3 to 6 words (i.e. 3- to 6-grams). The resulting key n-gram lists (for each of the three corpora compared to the other two, respectively) contain n-grams that appear significantly more frequently in one corpus (e.g. CHALC) than in another one (e.g. MICUSP_EL), and vice versa. Items are ordered according to their keyness value which is a measure for how 'key' an n-gram is in a corpus, i.e. how outstandingly frequent it is in one corpus compared to another one.

2.2 N-grams in Native and Non-native Speaker AAW

The first part of the analysis consisted of a careful examination and comparison of lists of frequent 3- and 4-grams in CHALC and MICUSP_EL. The 20 most common 3-grams and 4-grams in both corpora are displayed in Tables 1 and 2, together with their normalised frequencies (per million words). The first thing we observe is a considerable overlap of the items in the CHALC and MICUSP_EL columns in Table 1. Seven out of 20 3-grams appear in the top lists of both corpora. These are: *in order to*, *the fact that*, *as well as*, *the use of*, *part of the*, *one of the*, and *there is a* (set in small capitals in Table 1). These items are commonly used, both by native and non-native apprentice academic writers, to introduce evaluation or to structure the discourse. The 3-grams that only appear in the top-20 list of either CHALC or MICUSP_EL, fall into two groups: topic-related and non-topic-related 3-grams. Of the 3-grams that are related to the topics of the included papers (set in italics in Table 1), there are six in the CHALC list (e.g. *the present perfect*, *of functional grammar*, *in the clause*) and eight in the MICUSP_EL list (e.g. *in the novel*, *the united states*, *end of the*).[1]

Table 1: Top-20 3-grams in CHALC and MICUSP_EL

CHALC	hits (pmw)	MICUSP_EL	hits (pmw)
IN ORDER TO	761.1	IN ORDER TO	371.4
THE FACT THAT	615.7	ONE OF THE	321.2
the present perfect	426.6	THE FACT THAT	311.1
AS WELL AS	383	AS WELL AS	281
on the other	373.3	*the end of*	261
of the clause	353.9	*end of the*	230.8
the other hand	339.4	THERE IS A	225.8
THE USE OF	334.5	*at the end*	215.8
PART OF THE	329.7	THE USE OF	190.7
has to be	305.4	*the united states*	180.7
in the following	300.6	*in the novel*	160.6
of functional grammar	276.3	in which the	160.6
ONE OF THE	271.5	it is not	160.6
of the verb	261.8	there is no	160.6
in terms of	252.1	*in the first*	155.6
in contrast to	247.2	to be a	150.6
position of the	242.4	*nicaraguan sign language*	145.5
THERE IS A	242.4	PART OF THE	145.5
due to the	237.6	that it is	145.5
in the clause	237.6	*that she is*	145.5

The remaining items in the lists (seven out of 20 in CHALC and five out of 20 in MICUSP_EL) are not content-related but point towards differences between the students' academic writing styles. CHALC writers repeatedly use the 3-grams *on the other*, *the other hand* (→ *on the other hand*), *has to be*, *in the following*, *in terms of*, *in contrast to*, and *due to the* – all items that are also used by MICUSP_EL writers, though less frequently. On the other hand, the five 3-grams *in which the*, *it is not*, *there is no*, *to be a*, and *that it is* appear in the MICUSP_EL top-20 but not in the CHALC top-20 list. One thing we need to consider in this context are the altogether lower relative frequencies of occurrence of 3-grams in MICUSP_EL than in CHALC. This means that 3-grams that are among the top-20 in MICUSP_EL but not in CHALC may still be equally or more frequent in our German learner corpus, which is in fact the case for all of the five 3-grams mentioned above. If we also consider Hyland_Ling as a comparative resource, we find that the frequencies of n-grams in CHALC (i.e. the numbers of n-gram types) are on the whole very similar to those based on this expert academic writing corpus. This may indicate that our German advanced learners have a higher degree of academic writing competence and are more aware of the phraseological items commonly used in this genre than their native-speaker peers.

[1] A concordance analysis of *end of the* shows that this 3-gram is usually followed by either *novel*, *play*, or *story*.

A comparison of the lists in Table 1 with a top-20 3-gram list based on Hyland_Ling shows that all seven items that were found in CHALC and MICUSP_EL also occur in the Hyland_Ling list. Also, we find an overlap of three additional high-frequency items (*in terms of*, *on the other*, *the other hand*) with CHALC and of two additional items (*in which the*, *it is not*) with MICUSP_EL. The remaining eight 3-grams in the Hyland_Ling top-20 list do not occur in either of the AAW lists. Half of them are topic-related (e.g. *second language acquisition* and *the projected event*) but the other four are discourse-structuring devices (*the context of*, *the role of*, *some of the*, *in relation to*) that our apprentice academic writers do not use that frequently but which are important items in EAW.

Table 2: Top-20 4-grams in CHALC and MICUSP_EL

	hits		hits
CHALC	**(pmw)**	**MICUSP_EL**	**(pmw)**
ON THE OTHER HAND	339.4	THE END OF THE	185.7
AT THE END OF	174.5	AT THE END OF	170.6
in the case of	164.8	*in r j baumgardner*	100.4
with the help of	164.8	*r j baumgardner ed*	100.4
of the present perfect	150.3	ON THE OTHER HAND	90.3
THE END OF THE	145.4	*collected poetry and prose*	80.3
the specifier position of	135.7	it is clear that	80.3
quirk et al 1985	130.9	*of nicaraguan sign language*	80.3
to the fact that	130.9	one of the most	75.3
specifier position of the	126	the beginning of the	75.3
the fact that the	116.4	AT THE SAME TIME	70.3
the seven deadly sins	116.4	*the gospel of mark*	70.3
due to the fact	111.5	*the universal access plan*	70.3
of the english language	111.5	*the university of michigan*	70.3
the meaning of the	111.5	*the vicar of wakefield*	70.3
AT THE SAME TIME	106.7	as a result of	65.2
on the one hand	106.7	*in the united states*	65.2
the structure of the	97	it is important to	65.2
on the basis of	92.1	*new york w.w norton*	65.2
the subject of the	87.3	*the english language in*	65.2

A similar picture emerges if we consider the top-20 lists of 4-grams in Table 2. We observe an overlap of four items in both lists (*on the other hand*, *at the end of*, *the end of the*, *at the same time*), a large group of topic-related items that only occur in either CHALC (e.g. *the seven deadly sins*) or MICUSP_EL (e.g. *the gospel of mark*), and a few non-topic-related 4-grams that are either preferred by CHALC writers (e.g. *in the case of*, *the fact that the*, *on the one hand*) or by MICUSP_EL writers (e.g. *one of the most*, *as a result of*, *it is important to*). The same four items that appear in both lists in Table 2 also overlap with a Hyland_Ling top-20 4-gram list. Of the remaining 16 items in the Hyland_Ling list, eight are topic-related, three are shared with CHALC (*on the basis of*, *the fact that the*, *in the case of*), two are shared with MICUSP_EL (*as a result of*, *the beginning of the*), and three neither occur in the CHALC nor in the MICUSP_EL list (*in the context of*, *in terms of the*, *at the beginning of*), but are clearly useful phrases for academic writers.

As mentioned above in the data and methods section, the frequency n-gram analysis was complemented by a key n-gram extraction which was supposed to highlight phraseological items that are specific to only one corpus as compared to another one (rather than just retrieving n-grams that are more or less frequent in CHALC, MICUSP_EL and Hyland_Ling). Some important results from the key n-gram analyses are summarised in Tables 3 to 5 below. Table 3 lists the top-20 key n-grams in CHALC (with MICUSP_EL used as a reference corpus) and in MICUSP_EL (with CHALC used as a reference corpus). We note that, with only one exception in each of the two lists (*i.e. the* in CHALC and *it seems that* in MICUSP_EL, highlighted in small capitals), all key n-grams are content-related items, such as *quirk et al* or *of the verb* in CHALC and *indo pakistani english* or *of the narrative* in MICUSP_EL. This indicates that, when it comes to common non-topic-related phraseological items, CHALC and MICUSP_EL share a similar profile. Only when we scroll further down the key n-gram lists (to items with lower keyness values), do we find a few more items which are key in CHALC (compared to MICUSP_EL) and which tell us something about the phraseological profile of AAW by non-native speakers. These are: *with regard to*, *in contrast to*, *be regarded as*, *the help of*, *with the help*, *the other hand*, *on the other*, *e g in*, *in this context*, *regard to the*, *in order to*, *in the following*, *due to the fact*, and *on the other hand*. Key items in MICUSP_EL (compared to CHALC), though with lower keyness values than those displayed in Table 3, are: *way in which*, *is clear that*, *i did not*, *in this sense*, *a sense of*, and *that they have*.

Table 3: Top-20 key n-grams (span: 3-6) in CHALC and MICUSP_EL

key n-grams in CHALC (reference corpus: MICUSP_EL)	**keyness value**	**key n-grams in MICUSP_EL** (reference corpus: CHALC)	**keyness value**
halliday # #	810.4	the united states	183
the present perfect	466.4	nicaraguan sign language	176.5
of the clause	407.6	indo pakistani english	147
giegerich # #	317.4	university of michigan	145.6
harris # #	310.9	in the water	123.5
of functional grammar	303.1	gen # #	122
in the clause	267.8	j baumgardner ed	117.6
quirk et al	261.2	in genesis #	117.6
of the verb	231.5	poetry and prose	116.2
state of affairs	214.2	to the reader	111.8
the structure of	213.8	throughout the novel	110.3
the analysis of	208.9	the book of	108.8
halliday # # #	207.7	of the narrative	104.4
the speaker s	201.1	in the novel	96.5
et al #	198.5	of adult input	95.6
position of the	186.8	universal access plan	92.4
dik # #	186.8	IT SEEMS THAT	91.4
martin luther king	182.8	the pardoner is	91.2
I E THE	182.3	of nicaraguan sign	88.2
givón # #	177.6	vicar of wakefield	88.2

In order to see how the phraseological preferences of our CHALC and MICUSP_EL writers relate to expert academic writing, I also examined key n-gram lists for the two corpora with Hyland_Ling used as a reference corpus. The results, presented in Table 4, are very similar to those based on the CHALC/MICUSP_EL comparison (see Table 3). Except for one key 3-gram in each list (*has to be* in CHALC and *is able to* in MICUSP_EL), all key items tell us something about the topics of the papers included in the two corpora but not so much about the preferred discourse strategies of the apprentice academic writers. However, if we reverse the analysis and extract key n-gram lists from Hyland_Ling, using CHALC and MICUSP_EL as reference corpora, it becomes clear that there are indeed phraseological differences between expert and apprentice academic writing. Table 5 provides the resulting (filtered) lists of non-topic-related key n-grams in Hyland_Ling (or negative key n-grams in CHALC and MICUSP_EL). What is interesting here is that half of the key n-grams in the first list (reference corpus: CHALC, items set in small capitals) also appear in the second list (reference corpus: MICUSP_EL) which again points at similarities between native and non-native AAW as compared to EAW. That means that both CHALC and MICUSP_EL writers use the items in small capitals (e.g. *what was said* and *there has been*) significantly less often than Hyland_Ling writers. These items (together with some of the other items in Table 5, e.g. *in light of*, *is most likely* or *in terms of*) and their function and use would probably be worth focussing on in academic writing classes for native and non-native speakers of English.

Table 4: Top-20 key n-grams (span: 3-6) in CHALC and MICUSP_EL, compared with Hyland_Ling

key n-grams in CHALC (reference corpus: Hyland_Ling)	**keyness value**	**key n-grams in MICUSP_EL** (reference corpus: Hyland_Ling)	**keyness value**
halliday # #	791.9	in the novel	195.5
the present perfect	426.7	nicaraguan sign language	177.7
of the clause	398.5	that she is	173.3
giegerich # #	339.5	indo pakistani english	148.1
harris # #	332.5	of the novel	130.3
of functional grammar	324.1	in the water	124.4
quirk et al	279.4	gen # #	122.9
in the clause	252.3	in genesis #	118.5
halliday # # #	222.1	j baumgardner ed	118.5
HAS TO BE	207	poetry and prose	117
dik # #	199.7	throughout the novel	111.1
martin luther king	195.6	the book of	109.6
position of the	193.3	IS ABLE TO	105.4
chapter # #	192.8	of the narrative	105.2
givón # #	190	the gospel of	100.7
spencer # #	188.6	the story of	100.7
in chapter #	180.2	genesis # #	100.7
the specifier position	178.8	of women in	100.7
the mood element	171.8	of adult input	96.3
selkirk # #	163.4	universal access plan	93.3

Table 5: Non-topic related key n-grams (span: 3-6) in Hyland_Ling, compared with CHALC and MICUSP_EL

key n-grams in Hyland_Ling (reference corpus: CHALC)	**keyness value**	**key n-grams in Hyland_Ling** (reference corpus: MICUSP_EL)	**keyness value**
OF THE RESEARCH	112.9	IN THIS STUDY	108.7
OF THIS STUDY	110	OF THIS STUDY	105.2
IN THIS STUDY	84.5	WHAT WAS SAID	84.3
an attempt to	71.6	OF THE RESEARCH	79.9
were able to	68.3	the analysis of	76.5
is most likely	66.1	on the other	69.2
WHAT WAS SAID	64.5	the other hand	66.7
has been a	60.6	in terms of	65.5
I think that	60.6	the results of	62.3
THERE HAS BEEN	60.6	in this respect	58.3
in light of	60.6	THERE HAS BEEN	57
OF THIS RESEARCH	56.4	OF THIS RESEARCH	53.1

2.3 P-frames in Native and Non-native Speaker AAW

Let us now turn from n-grams to p-frames and look at pattern variability. Tables 6 and 7 present top-20 lists of high-frequency phrase-frames based on 3-grams (short '3-p-frames') and based on 4-grams (short '4-p-frames') extracted from CHALC and MICUSP_EL (with the floor set to three, meaning that only items that occur three or more times are included). The 'hits' column gives the total token frequency of all variants of the respective p-frame, e.g. 1,791 for the 3-p-frame *the * of*. The 'variants' column gives the number of variants (different types) a particular p-frame has, e.g. 209 for *the * of* which means that the empty slot (*) in the middle is filled by 209 different items in CHALC (e.g. *use*, *structure*, *meaning*, *end*). A high number in the variants column thus tells us something about the productivity of the p-frame. P-frames with high variant numbers (compared to their numbers of occurrence in a corpus), such as *to * the* or *is * to*, allow for more variation and are more productive than p-frames with low variant numbers, such as *in * to* or *that * is*.

If we compare the 3- and 4-p-frame lists based on CHALC and MICUSP_EL, we observe a great deal of overlap (14 out of 20 3-p-frames and nine out of 20 4-p-frames; indicated by small caps in Tables 6 and 7; most differences between the lists in Table 7 are topic-related, as the italicised p-frames indicate). That means that our native and non-native speaker apprentice academic writers commonly use many of the same p-frames, however with deviating frequencies. P-frame token numbers in MICUSP_EL are altogether lower than in CHALC, and again the CHALC-based figures are very close to those based on Hyland_Ling, which indicates that our non-native speaker apprentice academic writers seem to be able to master these items better than their native-speaker peers.

Table 6: Top-20 3-p-frames in CHALC and MICUSP_EL

	hits	variants		hits	variants
CHALC	**(tokens)**	**(types)**	**MICUSP_EL**	**(tokens)**	**(types)**
THE * OF	1791	209	THE * OF	1448	220
* OF THE	1191	157	* OF THE	940	137
OF THE *	1173	149	OF THE *	835	151
IN THE *	760	94	IN THE *	634	95
* IN THE	547	98	* IN THE	413	85
* TO THE	488	68	IT IS *	286	42
TO THE *	346	69	A * OF	228	39
* TO BE	321	30	TO * THE	201	42
IT IS *	297	38	* TO THE	176	37
A * OF	289	43	* IT IS	165	25
IN * TO	283	8	the * that	160	20
TO * THE	256	52	* TO BE	156	31
can be *	246	38	TO THE *	154	35
TO BE *	243	37	is * to	138	22
on the *	237	34	that * is	132	9
of a *	221	36	at the *	130	13
* IT IS	213	27	TO BE *	128	18
the * is	210	36	IN * TO	128	7
* that the	207	29	the * and	126	33
is * to	206	29	to * a	118	17

Table 7: Top-20 4-p-frames in CHALC and MICUSP_EL

	hits	variants		hits	variants
CHALC	**(tokens)**	**(types)**	**MICUSP_EL**	**(tokens)**	**(types)**
THE * OF THE	381	56	THE * OF THE	227	45
IN THE * OF	128	17	IN THE * OF	125	22
* THE FACT THAT	93	12	AT THE * OF	67	8
on the * hand	92	2	IN ORDER TO *	45	10
AT THE * OF	72	8	** the end of*	41	3
IN ORDER TO *	65	13	** end of the*	40	2
to the * of	63	16	THE END OF *	40	2
THE FACT THAT *	57	8	of the * of	32	9
** the present perfect*	53	5	it is * that	31	5
the * of a	51	12	* THE FACT THAT	27	5
With the * of	49	6	IT IS * TO	25	3
IT IS * TO	47	6	TO THE * THAT	24	5
** of the clause*	44	8	as a * of	23	4
*of the clause **	42	7	THE FACT THAT *	23	3
TO THE * THAT	40	3	The * in which	22	2
in * to the	39	5	as well as *	22	3
** position of the*	38	3	** the united states*	22	3
THE END OF *	38	3	that there is *	21	3
*the present perfect **	37	8	*j baumgardner ed **	20	2
*of the present **	36	2	a * of the	19	5

With respect to p-frame variation, the picture looks very similar for CHALC and MICUSP_EL. Not only are the numbers of variants for most of the high-frequency 3- and 4-p-frames in both corpora comparable, but CHALC and MICUSP_EL also to a large extent share the types of variants used in the p-frames, which points towards similar lexical preferences of the two groups of writers. Some p-frames, however, exhibit a lot of variation that is content-related. To give just two examples (one of the former and one of the latter type), *in * to* is (with only eight and seven variants) a very restricted or idiomatic 3-p-frame that shows the same variants in CHALC and MICUSP_EL: *order*, *contrast*, *addition*, *relation*, *reference*, *comparison*, and *regard* (plus *English* in CHALC).[2] Differences in lexical selection between CHALC and MICUSP_EL are exhibited by the 4-p-frame *in the * of*, for which the only items that appear in both lists are *case*, *context*, *form*, and *middle*. The remaining variants only appear either in CHALC (e.g. *analysis*, *teaching*, *grammar*, *meaning*) or in MICUSP_EL (e.g. *book*, *gospel*, *vicar*, *story*). These differences, however, are clearly due to the different topics covered in the papers included in the two corpora and hardly relate to the students' writing styles.

In a final analytic step, I examined Hyland_Ling-based 3- and 4-p-frame lists and compared them with the lists displayed in Tables 6 and 7 to see if there are any high-frequency p-frames in EAW that we do not find in AAW. The following items and their variants were found to be specific to Hyland_Ling and did not occur in the respective CHALC or MICUSP_EL lists: *in this ** (top variants: *case*, *study*, *way*, *article*, *respect*, *section*), *as a ** (top variants: *result*, *whole*, *way*, *means*), ** the other* (top variants: *on*, *of*, *and*), *on the * of* (top variants: *basis*, *use*, *part*, *distribution*), ** the context of* (top variants: *in*, *of*, *within*), *the * of this* (top variants: *results*, *purpose*, *implications*, *end*), and *in * of the* (top variants: *terms*, *spite*, *light*). These and other 3- and 4-p-frames that are common in Hyland_Ling but occur much less frequently in AAW (by native and non-native speakers) seem to be part of an expert academic 'phraseologicon' and probably deserve some special attention in EAP (English for Academic Purposes) teaching.

3. Summary and Conclusion: Nativeness and Academic Writing Expertise

In this paper, I set out to explore the phraseological profile of native and non-native speaker apprentice academic writing (AAW) and address the question whether, with respect to the use of phraseological items in written academic English, nativeness is an issue. I compared lists of frequent n-grams and p-frames (of different lengths) derived from three corpora (one each capturing native AAW, non-native AAW, and expert academic writing) with each other in order to see in what ways nativeness and expertise affect language patterning.

The n-gram and key n-gram analyses showed that there is a considerable overlap between the CHALC- and MICUSP_EL-based lists and that only a small number of n-grams refer to differences in academic writing styles among the two groups of students. Most differences could be explained on the basis of topic-related differences between the sets of papers included in CHALC and MICUSP_EL. We also saw that

[2] Largely the same variants are also found for this p-frame in a Hyland_Ling based list. An additional item here that does not occur in CHALC and MICUSP_EL is *opposition*.

native and non-native apprentice academic writers differ in similar ways from expert academic writers and that the CHALC writers are, in some respects, closer to the writers who contributed to Hyland_Ling than the MICUSP_EL writers. The p-frame analysis supported the findings from the n-gram explorations. Strong similarities were found between CHALC (NNS AAW) and MICUSP_EL (NS AAW), most of the differences we observed between the top p-frame lists were topic-related, and the comparison with Hyland_Ling highlighted some interesting differences between the use of p-frames by expert and apprentice academic writers.

These findings seem to indicate that, when we deal with advanced-level academic writing, we actually move beyond the native/non-native distinction and that, in this context, experience or *expertise* is a more important aspect to consider than nativeness (for more empirical evidence on this topic, see Römer 2009, and Wulff and Römer submitted). I agree with Swales (2004, 57), who sees a need to disentangle "communicative performance in research settings from mother-tongue status per se" and who uses an alternative distinction to NS vs. NNS, namely that of broadly English proficient (BEP) and narrowly English proficient (NEP) scholars. Another distinction used by Swales (2004, 56) that has been briefly referred to above and that is very much in line with our apprentice – expert divide is that between "senior" and "junior" scholars.

It appears that native and non-native apprentice academic writers develop their academic discourse competence in similar ways, and that native speakers also have to learn the language (and phraseology) of academic writing. The native academic writer does not seem to exist. Or, to quote Swales (2004, 52) once more, "[t]he difficulties typically experienced by NNS academics in writing English are (certain mechanics such as article usage aside) *au fond* pretty similar to those typically experienced by native speakers."[3] The fact that native and non-native apprentice academic writers lack very similar sets of expert academic English phraseological items in their papers indicates that both groups of students may need similar training or help with their academic writing on their way to becoming more proficient writers. For EAP teachers this implies, in Tribble's (2008, 307) words, that they "will be better served by using the notion of expertise [...] rather than the notion of the native-speaker." I would hence suggest that, in teaching academic writing, more emphasis be put on expertise than on nativeness and that writing instruction be based on samples of successful (or expert) writing by native or non-native speakers.

Acknowledgement

I wish to thank Matthew Brook O'Donnell for his help with the MICUSP_EL file extraction and conversion.

Works Cited

Bazerman, Charles. "Systems of genres and the enactment of social intentions". *Genre and the New Rhetoric*. Eds. Aviva Freedman and Peter Medway. London: Taylor and Francis, 1994. 79-101.

[3] I certainly experience these difficulties myself when I am asked to produce a piece of academic writing or give a talk in German, my native language. Since I am not used to using academic German, I find it much easier to function in English in academic contexts.

Bruce, Ian. *Academic Writing and Genre. A Systematic Analysis*. London: Continuum, 2008.

Fletcher, William H. *KfNgram*. Annapolis, MD: USNA, 2002-2007.

Flowerdew, John. "The non-Anglophone scholar on the periphery of scholarly publication". *AILA Review* 20 (2007): 14-27.

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

—. "The 'other' English: Thoughts on EAP and academic writing". *The European English Messenger* 15.2 (2006): 34-38.

Römer, Ute. "Learner language and the norms in native corpora and EFL teaching materials: A case study of English conditionals". *Anglistentag 2006 Halle. Proceedings*. Eds. Sabine Volk-Birke and Julia Lippert. Trier: Wissenschaftlicher Verlag Trier, 2007. 355-363.

—. "The inseparability of lexis and grammar: Corpus linguistic perspectives". *Annual Review of Cognitive Linguistics* 7 (2009), forthcoming.

Scott, Mike. *WordSmith Tools (Version 4.0).* Oxford: Oxford University Press, 2004.

— and Christopher Tribble. *Textual patterns. Key words and corpus analysis in language education*. Amsterdam: John Benjamins, 2006.

Sinclair, John McH. "Collocation: A progress report". *Language Topics. Essays in Honour of Michael Halliday*. Eds. Ross Steele and Terry Threadgold. Amsterdam: John Benjamins, 1987. 319-331.

—. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

—. "The search for units of meaning". *Textus* 9 (1996): 75-106.

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

Tribble, Christopher. "In this present paper... Some emerging norms in lingua franca English writing in the sciences?" *Proceedings from the 8th Teaching and Language Corpora* Conference *(TaLC8), Lisbon, Portugal, 3-6 July 2008*. Ed. Associação de Estudos e de Investigação Científica do ISLA-Lisboa, 2008. 307-309.

Wulff, Stefanie and Ute Römer. "Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive". *Corpora*, submitted.