View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Central Archive at the University of Reading

Formulaic sequences in native and non-native argumentative writing in German

Article

Published Version

Creative Commons: Attribution 3.0 (CC-BY)

Open Access

Jaworska, S., Krummes, C. and Ensslin, A. (2015) Formulaic sequences in native and non-native argumentative writing in German. International Journal of Corpus Linguistics, 20 (4). pp. 500-525. ISSN 1569-9811 doi: https://doi.org/10.1075/ijcl.20.4.04jaw Available at http://centaur.reading.ac.uk/40366/

It is advisable to refer to the publisher's version if you intend to cite from the work.

To link to this article DOI: http://dx.doi.org/10.1075/ijcl.20.4.04jaw

Publisher: John Benjamins Publishing Co.

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

University of Reading

Central Archive at the University of Reading

Reading's research outputs online

Formulaic sequences in native and non-native argumentative writing in German

Sylvia Jaworska, Cédric Krummes and Astrid Ensslin
University of Reading / Coventry University / Bangor University

The aim of this paper is to contribute to learner corpus research into formulaic language in native and non-native German. To this effect, a corpus of argumentative essays written by advanced British students of German (WHiG) was compared with a corpus of argumentative essays written by German native speakers (Falko-L1). A corpus-driven analysis reveals a larger number of 3-grams in WHiG than in Falko-L1, which suggests that British advanced learners of German are more likely to use formulaic language in argumentative writing than their native-speaker counterparts. Secondly, by classifying the formulaic sequences according to their functions, this study finds that native speakers of German prefer discourse-structuring devices to stance expressions, whilst British advanced learners display the opposite preferences. Thirdly, the results show that learners of German make greater use of macro-discourse-structuring devices and cautious language, whereas native speakers favour micro-discourse structuring devices and tend to use more direct language.

Keywords: formulaic language, n-grams, argumentative writing, German native speakers, advanced British learners of German

1. Introduction

The aim of this study is to compare the use of formulaic sequences identified in two corpora of German argumentative writings, the first corpus comprising argumentative essays produced by native speakers and the second one comprising comparable texts by advanced British learners of German. Following Biber et al. (2004), De Cock (1998) and Chen & Baker (2010), the analysis adopts a corpus-driven, frequency-based approach in order to investigate the frequency and the functions of 3-grams.

International Journal of Corpus Linguistics 20:4 (2015), **500–525**. DOI 10.1075/ijcl.20.4.04jaw
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company
This is an open access article licensed under the Creative Common Attribution 3.0 License (CC BY 3.0)

Quantitative and qualitative learner corpus research — as pioneered by the International Corpus of Learner English (ICLE) team led by Sylviane Granger (Granger 1998a, 1998b) — has provided invaluable insights into patterns of learner language, thus allowing for a systematic understanding of its lexical and grammatical idiosyncrasies. One aspect of learner corpus research which has attracted considerable research attention in recent years is formulaic language. This interest was prompted by Sinclair's (1991) pioneering work on corpus-driven lexicography, which highlighted the significance of the 'idiom principle' — a principle which provides evidence for the saliency of recurrent and semi-preconstructed chunks in language use (Granger 1998b). More recent studies into formulaic sequences have indeed demonstrated that formulaicity is a far more ubiquitous phenomenon than generative accounts lead us to believe (Altenberg 1998, Erman & Warren 2000). As studies in learner corpus research have shown, formulaic language seems to be frequently underused, overused or misused in learner language (Granger 1998b; De Cock 2000, 2004; Nesselhauf 2005), and hence it is increasingly seen as the element distinguishing advanced L2 learners from native speakers.

Despite the rapid development of learner corpus research, most studies have, to date, focused predominantly on English as a Second or Foreign Language (L2 English hereafter). We now have a solid body of results demonstrating typical lexical and structural patterns of L2 English acquired in a variety of linguistic and cultural contexts. In recent years, learner corpora for languages other than L2 English have been compiled, mostly for L2 Spanish and L2 French. Yet, published research in this area is, as compared with L2 English, still scarce.

As a result, there is little corpus evidence of typical lexico-grammatical patterns of other L2s. German is a good example. Most research on German as a Second or Foreign Language (L2 German hereafter) has been preoccupied with cognitive mechanisms or selected syntactical and morphological phenomena underlying the acquisition process and has been based on traditional manually-encoded error analysis (Wend 1998). While this research provided valuable insights into the development of L2 German, the findings were often based on small data sets, collected in an L2 environment (i.e. in German-speaking countries), making generalisations about other, instructed learning L1 environments problematic. At present, there is little empirical evidence for the existence of typical patterns of overuse or underuse as produced in instructed learning settings, which is a serious gap, given the importance of L2 German in language classrooms outside German-speaking countries. German is, for example in Europe, the second or the third foreign language after English (see Education, Audiovisual and Culture Executive Agency 2008) and the vast majority of learners acquire German in institutional contexts, such as schools and universities. Neuner (2004) estimates the total number of German language learners in the world to be between 15 and

20 million, of whom three-quarters are based in Europe, including Central and Eastern Europe.

However, with a few exceptions (Belz 2004, Möllering 2004, Maden-Weinberger 2008), there has been, up to now, very little systematic corpus-based research into the lexico-grammatical patterns of L2 German in linguistic and cultural contexts outside German-speaking countries. This is partially due to a lack of systematically compiled German learner corpora. It was only recently that this shortcoming has been recognised. In this respect, the Falko project led by Anke Lüdeling at Humboldt University Berlin is pioneering, as it is the first large corpus of advanced learner German with data collected from learners of various L1 (49 languages as L1 in total) (Lüdeling et al. 2008).

Whereas previous learner corpus research has concentrated on formulaicity in L1 and L2 English, this paper examines the use of recurrent sequences produced by native and non-native speakers of German. The data under scrutiny consists of two corpora: the first being a corpus of argumentative essays written by advanced British students of German (WHiG) and the second being a corpus of comparable argumentative essays written by German native speakers (Falko-L1). Both corpora are part of the aforementioned Falko project. Argumentative essays are worthwhile to investigate as they form an inherent part of any advanced foreign language learning programme.

Drawing on previous corpus-driven and frequency-based research on formulaic language (De Cock 2004, Biber et al. 2004, Juknevičienė 2009, Chen & Baker 2010), our aim is to identify quantitative and qualitative similarities and differences in the use of formulaicity by native and advanced non-native speakers of German. In so doing, this research attempts to contribute to the slowly growing body of learner corpus research on formulaic language in L1 and L2 varieties other than English.

The issue of comparing learner language with native output has been a matter of much debate in research literature. Given that the percentage of L2 learners who achieve native-speaker competence is nil to 5% (Han 2011), some researchers argue that using the target language as a benchmark for comparison is not adequate. Following Kramsch (1997) and Cook (1999), we agree that, from a pedagogical point of view, using native speakers as role models is a problematic issue and that native speaker norms are not necessarily appropriate in an L2 context. However, at the same time, it cannot be ignored that the native norm is something many learners aspire to, and it would be equally unfair to disrespect their ambition even if the target may be achievable for only a few (Timmis 2002).

We begin our investigation by providing an overview of previous research into formulaicity in learner language. We focus in particular on studies that have been based on learner corpora. This is then followed by a methodological section

presenting the data collection procedures and research methods adopted in this study. Our results shed important light on the specific linguistic and discursive challenges facing advanced British learners of German. In our concluding section we offer some suggestions as to how the research findings can be used for the development of teaching and learning materials for L2 German, pitched specifically at advanced British learners.

2. Learner corpus research into formulaic language

The term 'formulaic language' is normally used to denote multi-word sequences in language that appear to be stored in the mind as holistic units and retrieved as complete chunks from memory (Wray 2002: 9). Formulaicity has been described by a wide range of terms including 'collocations', 'recurrent word combinations', 'clusters', 'n-grams' and 'lexical bundles'. In her extensive overview of formulaic language, Wray (2002) identifies more than 60 various labels. In learner corpus research, the terminology depends largely on the methodological procedures adopted to study formulaicity. For example, studies that favour a category-based approach tend to use the term 'prefabricated pattern' (Granger 1998b) or 'collocation' (Nesselhauf 2005, Laufer & Waldman 2011), while corpus-driven frequency-based research adopts mostly the notion of 'lexical bundle' associated with work by Biber et al. (2004) or the term 'n-gram' taken from computational linguistics (O'Donnell et al. 2013). As highlighted recently by McEnery & Hardie (2012: 110), the terms 'lexical bundle' and 'n-gram' are methodologically and technically the same, although the former is more often associated with retrieval procedures and the latter with a functional and structural interpretation of n-grams. Since this study follows the corpus-driven frequency-based methodology and involves an automatic retrieval of recurring sequences of *n* words, we will use the term 'n-gram' to refer to our procedures and results. When referring to formulaicity in general, we will adopt, following Wray (2002), the most neutral term 'formulaic language'.

Traditionally, linguists have employed the term 'formulaic language' to describe fixed phrases such as idioms, proverbs and sayings that are rare in language use (Granger & Paquot 2008). Recent corpus-driven evidence has demonstrated, however, that formulaicity is a much more salient and ubiquitous linguistic phenomenon than previously thought and includes sequences ranging from completely fixed strings of words (idioms and fixed expressions such as e.g. *happy birthday*) to far more flexible structures with a greater morphological and/or lexical and syntactical variability, for example collocations such as *exert/wield influence* or phrase-frames such as *if you look at* * or *I don't know what* *. Erman & Warren (2000) estimate that nearly 60% of spoken English and more than 50% of

written English consists of such diverse formulaic sequences. The frequent use of formulaic sequences is driven by the principle of economy of effort. Wray (2002) suggests that because formulaic sequences are stored and retrieved as single units, their use ensures a considerable reduction of the processing time needed to decode and encode information. In doing so, they guarantee communicative efficiency, fluency and smoothness (Kuiper & Haggo 1984). They are not necessarily complete or well-defined linguistic structures but rather lexico-grammatical fragments, which "function as basic building blocks of discourse" (Biber et al. 2004: 1) ensuring cohesion and coherence of speech and writing.

Parallel to the corpus research on formulaicity in L1 English, there have been a number of learner corpus studies into formulaic sequences in L2 English. In learner corpus research on formulaic language, two methods are commonly used to detect formulaic sequences. Some researchers favour a category-based approach, which identifies formulaic sequences on the basis of linguistic categories set a priori, for example, combinations of selected parts of speech, commonly referred to as collocations. A corpus is then searched for instances of the predefined structures. Studies following the category-based approach suggest that L2 learners tend to underuse native-like expressions as compared with native speakers. For example, Granger (1998b) has shown that L2 learners tend to underuse native-like expressions, yet overuse those word pairs which have direct L1 equivalents. They often produce overlaps, i.e. sequences blending native-like with non-native strings of words and tend to have a smaller repertoire of collocations, with restricted collocations being particularly problematic items to learn (Nesselhauf 2005), even at advanced levels (Laufer & Waldman 2011).

The second approach, referred to as distributional, frequency-based or lexical bundle approach is associated with work by Altenberg (1998), Biber et al. (2004) and De Cock (2000) and is based on the automatic retrieval of n-grams, that is, recurrent strings of two, three or more words. Although sequences extracted in this way are not necessarily complete structural units but rather constitute incomplete lexico-grammatical fragments, they are recognised as important routinised building blocks of discourse (Altenberg 1998, Biber at al. 2004), markers of L2 fluency (Hyland 2008a) and a possible quantitative measure of L2 lexical development (Groom 2009). For example, by analysing quantitatively a larger corpus of native and non-native speech and writing, De Cock (2000, 2004) demonstrates that learners tend to overuse recurrent word combinations in both written and spoken registers and that there is a less marked difference between writing and speech in L2 output, pointing to a stylistic deficiency.

In comparing essays written in English by native speakers and Lithuanian speakers, Juknevičienė's (2009) findings confirm that L2 speakers use more lexical bundles than L1 speakers and rely more on "frequent repetition of "safe" phrases"

(Juknevičienė 2009: 65). More precisely, her study reveals that L1 speakers use fewer different lexical bundles (types) and a lower number of lexical bundles (tokens) than L2 speakers. Juknevičienė (2009) further observes that, structurally and functionally, her L2 corpora reveal lexical bundles which are more typical of spoken than written English, whereas the L1 corpus bears stronger resemblances to expert academic writing.

In a similar vein, Chen & Baker (2010) retrieve 4-grams from three corpora of academic writing: the first including native expert writing, the second native student writing and the third L2 student writing (L1 Chinese). In contrast to previous research, Chen & Baker (2010) have demonstrated that non-native speakers produce fewer formulaic sequences than native speakers. The authors also observe that both native and non-native student essays exhibit features that distinguish them from professional native expert prose. L1 English student essays show more "control of cautious language" (Chen & Baker 2010: 44) by using significantly more hedges and other low-modality formulations, such as the sequence *is likely to be*. This kind of language is not present in the L2 essays of advanced Chinese EFL learners, which in contrast tend to be "stylistically more verbose" (Chen & Baker 2010: 43) using repetitions or tautologies — a set of features which many scholars see as a common trait of L2 academic writing.

The results emerging from the above studies do seem to be somewhat inconsistent. While studies using the category-based approach point to the pattern of underuse, the frequency-based methodology delivers results indicating overuse. Undoubtedly, the method used will have an impact on results. Both approaches yield empirical quantitative insights. The category-based approach is a corpus-based type of research and as such interrogates a corpus for the existence of selected patterns (Tognini-Bonelli 2001). It offers insights into the use of the selected category and any findings can be generalised to that selected pattern only, for example Verb-Noun collocations. Corpus-driven research, on the other hand, takes the available data as a whole and claims to use very little theoretical presupposition about grammar and lexis. For this reason, some prefer the corpus-driven methodology over the corpus-based because of its inductive, bottom-up nature, which purportedly reduces biases. While in theory the whole data should be interrogated, in practice this is rarely the case, as corpus-driven research often applies frequency as a filter to eliminate some sections of the data (Groom 2009). Also, the bias- and theory-free claim does not necessarily hold. Corpus-driven results often yield thousands of patterns that are subsequently categorised by referring to pre-existing theoretical models of grammar and lexis. Hence, both approaches have certain advantages and limitations and at times, the difference between the two seems to be blurred (McEnery & Hardie 2012). For investigating formulaic patterns, the definition of what constitutes formulaicity should guide the choice

of method. If formulaic sequences are defined on the basis of pre-existing lexico-grammatical categories such as parts of speech, then the corpus-based approach is more appropriate. If formulaicity is understood as sequences of recurrent word-combinations, then a corpus-driven methodology seems more suited.

For the purpose of this research, formulaic sequences are defined as a series of automatically retrieved sequences of *n* words identified on the basis of frequency that have customary pragmatic and/or discourse functions (Biber at al. 2004). Hence, a corpus-driven design was adopted with the aim of examining the distribution and functions of the most frequent 3-grams produced by native and non-native speakers of German in argumentative writing. To our knowledge, there have been no published studies that examine recurrent word sequences in German, native or non-native, and this study is the first of this kind. The main research questions that this study addresses are:

i. How many 3-grams (types and tokens) are found in native and non-native corpora?
ii. What are the most frequent 3-grams and how do the corpora differ?
iii. What are the functions of the most frequent 3-grams?
iv. Can any functional differences be detected between the use of 3-grams by native and non-native speakers of German?

An in-depth quantitative and qualitative examination of 3-grams that are attested in the German native speaker corpus but absent or modified in the corpus of British learners of German will contribute to a better understanding of lexico-grammatical patterns used by advanced British learners of German compared to native speakers. Our results also provide a number of suggestions for the design of evidence-based teaching and learning materials for advanced Anglophone learners of German.

3. Corpora and methodology

In this section, we discuss the main steps that were involved in compiling our unique corpora. This is followed by the discussion of the analytical procedures used to retrieve and analyse 3-grams.

3.1 Data collection

The data for the present study comes from two German corpora, Falko-L1 and WHiG, from which formulaic sequences were automatically retrieved (see Section 3.2). Both corpora are part of the parent project Falko (Lüdeling 2011).

The native-speaker corpus Falko-L1 consists of 116 essays and 77,357 tokens. The L1 German speakers are secondary school leavers from in or around Berlin. The learner corpus WHiG ("What's Hard in German?") consists of 173 essay files and 90,883 tokens. The essays were composed by British students from six universities in England and Wales who were — at the time of the data collection (February 2010 to November 2011) — undergraduates studying German as Single Honours or in combination with another subject. Participants were given a choice of four topics[1] and 90 minutes to write approximately 500 words without the aid of any German grammar spellchecker. Table 1 provides a breakdown of the number of tokens and types in Falko-L1 and WHiG.

Table 1. Composition of Falko-L1 and WHiG

	Falko-L1	**WHiG**
number of essays	116	173
mean tokens per essay	666.87	525.33
tokens	77,357	90,883
types	9,996	7,463
TTR	12.99	8.23
standardised TTR	44.26	43.70

Because the parent project Falko focuses on advanced learners of German, only data from learners who have reached at least the B2 level on the Common European Framework of Reference for Languages (CEFR) scale was considered, as B2 is the required German L2 entry level for German universities. To determine the participants' level of proficiency, WHiG respondents had to complete a C-test (Raatz & Klein-Braley 1982), a statistically approved type of cloze test widely used in the context of German as a Foreign Language to determine the proficiency level. A participant achieving a C-test score of between 60 and 79, for instance, was assigned the level B2, whereas C1 was between 80 and 89, and C2 speakers had a score of 90 or above (see Table 2).

1. The four topics were: (i) Kriminalität zahlt sich nicht aus "Crime does not pay", (ii) Die meisten Universitätsabschlüsse sind nicht praxisorientiert und bereiten die Studenten nicht auf die wirkliche Welt vor. Sie sind deswegen von geringem Wert "Most university degrees are not hands-on and do not prepare students for the real world. They are therefore of little value", (iii) Die finanzielle Entlohnung eines Menschen sollte dem Beitrag entsprechen, den er/sie für die Gesellschaft geleistet hat "A person's financial reward should be commensurated with his/her contribution to society", and (iv) Der Feminismus hat den Interessen der Frauen mehr geschadet als genützt "Feminism has done more harm to the cause of women than good".

Table 2. C-test scores and their CEFR level equivalent

C-test Score	CEFR level
90–100	C2: Mastery
80–89	C1: Effective Operational Proficiency
60–79	B1: Vantage
40–59	B2: Threshold
30–39	A2: Waystage
0–29	A1: Breakthrough

The present study is based on an analysis of argumentative essays ("Erörterungen"). These are normally not recognised as fully fledged academic texts, since they lack references or a rigid mesostructure. They can contain elements of everyday spoken language such as shorter sentences, more paratactic and fewer hypotactic sentences, conditional constructions with the *würden* auxiliary and personal statements in the first person singular *ich* (Sieber 1998: 194). At the same time, argumentative essays share a number of linguistic and conceptual features with the style of German academic writing such as: *Nominalstil* "nominalised style", longer argumentative structures and critical analysis (cf. Fix 2008). Most importantly, however, argumentative essays function as a pre-stage to academic writing and, particularly from a lexical point of view, can be useful in assisting novice writers in the development of academic literacy.

3.2 Procedures of data analysis

Two word frequency lists were retrieved from Falko-L1 and WHiG by using *WordSmith Tools* version 5 (Scott 2008). The software search was set to retrieve only 3-grams and 4-grams that occurred at least 5 times in at least 3 texts or 2.5% of the texts in the sample. We agree with Biber et al. (2004: 376) that the parameters set to identify lexical bundles are "somewhat arbitrary". Earlier research that utilised large corpora including millions of tokens suggests a cut-off of 40 times per million words (Biber et al. 2004) or an occurrence in at least 10% of texts (e.g. Hyland 2008b). Our data sets are too small (below 100,000 tokens) to adopt a cut-off based on *x* occurrences per million words, as this would inflate the rate of occurrences (Biber & Barbieri 2007); 10% of all texts as a cut-off for our corpora would produce a very small number of ngrams. We felt that having the cut-off point set at 5 times in at least 3 texts offers a sufficient measure to guard against idiosyncratic uses in our smaller specialised corpora.

After the retrieval of 3-grams and 4-grams, it became apparent that the number of 4-grams was too small (see Table 3) to offer any insights into lexico-grammatical patterns of learner and native German and most of them were sequences copied

directly from the essay topics. This is an interesting result given that previous corpus-driven research sees 4-grams as the most productive units to examine (Biber & Barbieri 2007, Hyland 2008a). However, this may only be suitable for analytic languages such as English that exhibit a reduced inflectional morphology and a more rigid syntactic order. German, which is a synthetic language, allows a greater flexibility of syntax. Intuitively, this may have an effect on the retrieval of fewer 4-grams when using the corpus-driven approach, which proceeds in a linear manner without taking into account syntactic variability. For example, the 4-gram *there are many examples*, which is represented by only one syntactic structure in English, can have 3 syntactic variants in German declarative sentences ("es gibt viele Beispiele", "viele Beispiele gibt es", "[…], gibt es viele Beispiele"). Because the variants may occur with different frequencies, 4-grams may not capture all of them. The fact that more 3-grams were obtained could suggest that sequences of three items reflect formulaicity better in German. As can be seen below, 3-grams also seem to capture syntactic variability adequately, because the items which they contain are often high frequency words. The decision was thus taken to examine in more depth 3-grams only.

Table 3. 3-grams and 4-grams with two cut-off points

	Falko-L1		**WHiG**	
	types	**tokens**	**types**	**tokens**
3-grams	202	1,499	457	4,197
4-grams	32	242	80	726

Following the taxonomy proposed by Biber et al. (2004), the formulaic sequences were then categorised according to their function. The functions included: (i) reference markers, that is, sequences naming physical and abstract objects, spatial and temporal references, qualities, and quantities; (ii) discourse-structuring markers understood as bundles that organise the text; (iii) stance markers that express the writer's attitude or an evaluation of a proposition in terms of certainty or uncertainty. In addition, the 3-grams directly copied from the essay topics were described as a separate "topic" category. A smaller number of formulaic sequences had no distinctive function or meaning and were labelled as "unclassified".

4. Results

This section reports on the main quantitative and qualitative results obtained in the present study and it is divided into two subsections. Whereas the first part focuses on the most frequent 3-grams attested in both corpora, the second part offers qualitative insights into the functions of the retrieved sequences.

4.1 Quantitative analysis

Table 4 shows a breakdown of the 3-grams (by types and tokens) found in the two corpora. A distinction is made between 3-grams that were directly copied from the essay topics (e.g. *für die gesellschaft* "for society", *kriminalität zahlt sich* "crime [does not] pay") and those that were not (marked as "topic" and "non-topic" in Table 4). To ensure that all 3-grams are captured, all words were treated as lowercase and are represented as such in the tables.

As can be seen in Table 4, the WHiG corpus contains a much higher frequency of 3-grams. Learners are on average three times more likely to use these sequences, with regards to both tokens and types.

Table 4. 3-Grams (cut off point 5; normalisation per 10,000 tokens)

		Falko-L1		WHiG	
		raw	**norm.**	**raw**	**norm.**
non-topic	tokens	1,033	133.54	2,946	324.15
	types	139	139.05	334	447.54
topic	tokens	466	8.14	1,246	137.09
	types	63	63.02	123	164.81

This seems to confirm the results obtained in previous studies (De Cock 2000, 2004; Juknevičienė 2009) that learners tend to rely more on formulaic sequences than native speakers. However, the results contrast with Chen & Baker's (2010: 33) findings, where it is the expert English-L1 writers (FLOB-J corpus) that produce more types and more tokens. Hypothetically, this discrepancy may be due to the fact that Chinese learners of English (as reported in Chen & Baker 2010) are facing a far greater linguistic difference between L1 and the target language and are therefore more likely to use a smaller yet overused set of formulaic sequences. Moreover, the texts examined in their study were examples of complex academic register such as academic essays and published academic work that possibly relies on fewer lexical bundles than the type of argumentative essays analysed in the present study.

Table 5 shows the twenty most frequent 3-grams in Falko-L1 and WHiG. As can be seen, many of the most frequent 3-grams in both corpora are sequences copied directly from the essay topics (in italics). If we omit them, it becomes apparent that there are only three 3-grams in the lists that are shared by both groups: *meiner meinung nach* "in my opinion", *der meinung dass* "of the opinion that", *man sagen dass* "one say that". Otherwise, both sets include different combinations. For example, the most frequent 3-grams in Falko-L1 are prepositional phrases including prepositions requiring the dative and accusative case. These are less frequent in the WHiG data. The use of prepositions with the dative and accusative case

Table 5. The 20 most frequent 3-grams in Falko-L1 and WHiG

Falko-L1	Raw freq.	Norm. freq.	WHiG	Raw freq.	Norm. freq.
meiner meinung nach	42	542.93	an der universität	45	495.14
beitrag für die	28	361.95	*kriminalität sich nicht*	40	440.12
in den letzten	21	271.46	meiner meinung nach	38	418.12
dass sich kriminalität	18	232.68	in der gesellschaft	37	407.11
in den meisten	17	219.76	an der uni	35	385.11
in der gesellschaft	17	219.76	der meinung dass	35	385.11
in unserer gesellschaft	17	219.76	*dass der feminismus*	32	352.10
auf jeden fall	16	206.83	es gibt auch	32	352.10
der meinung dass	16	206.83	*feminismus den interessen*	31	341.10
ich denke dass	15	193.91	*sich nicht auszahlt*	30	330.09
sich kriminalität nicht	15	193.91	der anderen seite	28	308.09
den ganzen tag	14	180.98	*der feminismus den*	28	308.09
der feminismus den	14	180.98	*der wirklichen welt*	28	308.09
den letzten jahren	13	168.05	*nicht praxisorientiert sind*	27	297.08
die frage ob	13	168.05	*dass kriminalität sich*	26	286.08
in der heutigen	13	168.05	ist es nicht	26	286.08
wie zum beispiel	13	168.05	man sagen dass	26	286.08
kriminalität nicht auszahlt	12	155.12	auf der anderen	25	275.08
rolle der frau	12	155.12	es gibt viele	25	275.08
dass der feminismus	11	142.20	zu sagen dass	25	275.08

is one of the stumbling blocks for learners of German and the smaller number of such constructions in WHiG might suggest that learners tend to avoid them. In contrast, the WHiG data contains more clause fragments, of which the most frequent are combinations containing the existential *es gibt* "there is/are". These are: *es gibt auch* "there is/are also", *es gibt viele* "there are many" and its syntactical variant *gibt es viele* "are there many". Such constructions are not very frequent in the counterpart corpus. For example, in FALKO-L1, there are only seven types of 3-grams (39 tokens) with the existential *es gibt/gibt es*, whereas there are 34 types in WHiG amounting to 300 tokens (see Table 6).

The higher frequency of 3-grams with *es gibt* and *gibt es* in the WHiG corpus suggests that texts produced by British learners rely heavily on existentials. The combinations also include lexical elements not observed in the native data, for example *jedoch gibt es* "however, there is". To further test this hypothesis, we also examined the use of another typical existential of German in both corpora, namely *es ist* "it is" and its syntactical variant *ist es* "is it". The analysis too demonstrates the overuse pattern of this existential. Whereas there are only four 3-grams with

Table 6. 3-grams with the existential *es gibt/gibt es*

WHiG	Raw freq.	Norm. Freq.	Falko-L1		Norm. Freq.
es gibt auch	32	352.1	es gibt viele	8	103.4
es gibt viele	25	275.08	natürlich gibt es	6	77.56
gibt es viele	19	209.06	dennoch gibt es	5	64.64
aber es gibt	14	154.04	doch es gibt	5	64.64
gibt es immer	14	154.04	es gibt auch	5	64.64
und es gibt	14	154.04	gibt es noch	5	64.64
es gibt aber	11	121.03	und es gibt	5	64.64
es gibt noch	11	121.03			
es ist auch	11	121.03			
gibt es ein	9	99.028			
heutzutage gibt es	9	99.028			
gibt es auch	8	88.025			
gibt es eine	8	88.025			
gibt es die	7	77.022			
gibt es noch	7	77.022			
gibt es so	7	77.022			
es gibt ein	7	77.022			
es gibt viel	6	66.019			
gibt auch viele	6	66.019			
es gibt die	5	55.016			
es gibt eine	5	55.016			
es gibt jedoch	5	55.016			
es gibt natürlich	5	55.016			
es gibt nicht	5	55.016			
auch gibt es	5	55.016			
deutschland gibt es	5	55.016			
es gibt die	5	55.016			
es gibt eine	5	55.016			
es gibt jedoch	5	55.016			
es gibt natürlich	5	55.016			
es gibt nicht	5	55.016			
gibt es viel	5	55.016			
jedoch gibt es	5	55.016			
seite gibt es	5	55.016			

es ist/ist es in Falko-L1, we find 31 such sequences in WHiG (see Table 7). Overall, the learners in the context under study are 13 times more likely to use a combination with *es ist/ist es* than native speakers. The data also reveals a variety of items

Table 7. 3-grams with the existential *es ist/ist es*

WHiG	Raw freq.	Norm. Freq.	Falko-L1	Raw freq.	Norm. Freq.
es ist nicht	18	198.1	so ist es	7	90.49
und es ist	18	198.1	es ist also	6	77.56
aber es ist	16	176.1	es ist ein	5	64.64
ist es klar	14	154	ist es auch	5	64.64
es ist auch	11	121			
es ist ein	11	121			
es ist eine	10	110			
deshalb ist es	10	110			
es ist klar	9	99.03			
deswegen ist es	9	99.03			
ist es wichtig	9	99.03			
ist klar dass	9	99.03			
vielleicht ist es	9	99.03			
es ist oft	8	88.03			
heutzutage ist es	8	88.03			
ist es aber	8	88.03			
ist es sehr	8	88.03			
jedoch ist es	7	77.02			
es ist ganz	6	66.02			
es ist sehr	6	66.02			
also ist es	6	66.02			
ist es möglich	6	66.02			
es ist aber	5	55.02			
es ist wichtig	5	55.02			
heute ist es	5	55.02			
ist es auch	5	55.02			
ist es eine	5	55.02			
ist es ganz	5	55.02			
ist es oft	5	55.02			
ist es schwer	5	55.02			
ist für viele	5	55.02			

in the left or right co-text of the existential including adjectives, conjunctions and adverbs — something which is not matched by the Falko-L1 data.

All in all, the quantitative analysis reveals that learners use more formulaic sequences than native speakers. However, if we take a closer look at the syntactical constructions on which some of the most frequent combinations are based,

it becomes obvious that many of them are composed of repetitive and simple syntactical fragments, mainly existentials.

4.2 Functions of formulaic sequences

Table 8 and 9 list all 3-gram types and tokens categorised according to their main function. As some sequences may have more than one function, concordance lines were carefully checked in order to identify the dominant function of each sequence — a procedure also applied in previous research (Biber et al. 2004, Chen & Baker 2010).

Table 8. Functions of 3-gram (types) in Falko-L1 and WHiG

	Falko-L1		WHiG	
	Raw Freq.	**%**	**Raw Freq.**	**%**
discourse-structuring	60	29.70	114	24.95
referential	36	17.82	61	13.35
stance	34	16.83	124	27.13
topic	63	31.19	123	29.91
unclassified	9	4.46	35	7.66

Table 9. Functions of 3-grams (tokens) in Falko-L1 and WHiG

	Falko-L1		WHiG	
	Raw Freq.	**%**	**Raw Freq.**	**%**
discourse-structuring	397	26.48	1,018	24.28
referential	296	19.75	641	15.29
stance	286	19.08	1,080	25.76
topic	466	31.09	1,246	29.72
unclassified	54	3.60	207	4.94

As shown in Table 8, "topic" sequences are the largest category and have a similar proportion in both data sets suggesting that both groups rely on the sequences used in the topic prompts. Interestingly, more types of discourse-structuring devices are found in Falko-L1 (29.70%) than in WHiG (24.95%), whereas stance expressions are more common in WHiG (27.13%) than in Falko-L1 (16.83%). The difference in the use of referentials is much smaller, though there is a slightly higher proportion in Falko-L1 than in WHiG. In the WHiG data, we have more types of 3-grams that could not be assigned any of the functions because they consisted of grammatical words only. These include sequences such as: *sie in der* "they/she in the", *ist und es* "is and it" or *ist weil es* "is because it".

A token distribution yielded similar results (see Table 9 above). Still, there is a higher proportion of discourse structuring-devices in Falko-L1. These sequences constitute the second largest category after topic-related expressions. In contrast, the WHiG data demonstrates a higher proportion of stance expressions, which in this data set rank second.

A chi-square test for both types and tokens shows that there is a significant difference in terms of the functional distribution between the two data sets, for types: $x^2(4,\ N=659)=12.3655$, $p=0.01483$, and for tokens: $x^2(4,\ N=5{,}691)=40.9911$, $p=2.699\text{e-}08$.

It can, therefore, be concluded that underuse of discourse-structuring devices and overuse of stance expressions seem to be a typical feature of argumentative writing composed by advanced British learners of German as compared to the native counterparts.

A chi-square test comparing topic 3-grams and all other 3-grams in both corpora shows no significant results, for types $x^2(1, N=659)=1.2628$, $p=0.2611$, and for tokens: $x^2(1, N=5{,}691)=0.9769$, $p=0.323$.

In a next step, we examined the different types of the most frequent 3-gram in the following three categories: discourse-structuring devices, stance and reference expressions (see Tables 10, 11 and 12). As most of the topic sequences were

Table 10. The 20 most frequent discourse-structuring devices (raw frequencies)

Falko-L1	WHiG
die frage ob (13)	es gibt auch (32)
wie zum beispiel (13)	der anderen seite (28)
sich die frage (11)	auf der anderen (25)
auch wenn sie (10)	es gibt viele (25)
den meisten fällen (10)	in bezug auf (22)
doch was ist (10)	wenn man ein (22)
wenn man sich (10)	gibt es viele (19)
auf der anderen (9)	wenn man eine (19)
es sich um (9)	wie zum beispiel (15)
gibt es auch (9)	aber es gibt (14)
ob es sich (9)	gibt es immer (14)
zu tun haben (9)	im vergleich zu (14)
der anderen seite (8)	und es gibt (14)
es gibt viele (8)	zum beispiel wenn (14)
alles in allem (7)	in diesem aufsatz (13)
an dieser stelle (7)	wenn man etwas (13)
es stellt sich (7)	auch wenn man (12)
frage ob sich (7)	dass wenn man (12)
zu diesem thema (7)	man zum beispiel (12)
zum beispiel die (7)	und wenn man (12)

combinations of words appearing in the titles of the essays, these were excluded from further analysis. The unclassified formulaic sequences were not considered either.

As can be seen in Table 10, most of the discourse structuring devices in WHiG contain the existential *es gibt* "there is" or the clause fragment *wenn man* "if one". Most of them tend to be topic or argument initiators: they structure larger units of text or text as a whole (macro-discourse). The cluster *in diesem aufsatz* "in this essay" is a paramount example of a device structuring the macro-text, the essay as a whole. There are in total 3 types of combinations with *Aufsatz* "essay" in WHiG including *dieser aufsatz wird* "this essay will" and *aufsatz werde ich* "essay will I", which is part of a larger sequence *in diesem essay werde ich* "in this essay I will". Such devices are, interestingly, absent in Falko-L1. In the latter, we find many instances of discourse-structuring combinations whose function is to ensure cohesion and coherence at the micro-level of discourse such as question markers (*die frage ob* "the question whether"), summarisers (*alles in allem* "all in all"), contrast markers (*doch was ist* "however what is"; *auf der anderen/der anderen seite* "on the other hand"), exemplifiers (*wie zum beispiel* "as for example") and text-deictic expressions such as *an dieser stelle* ("at this point"). Apart from the use of the contrast marker *(auf) der anderen seite* "on the other hand" and exemplifiers containing the phrase *zum beispiel* "for example", we do not find many instances of such devices in the WHiG

Table 11. The 20 most frequent reference-marking sequences (raw frequencies)

Falko-L1	WHiG
in den letzten (21)	an der universität (45)
in den meisten (17)	in der gesellschaft (37)
in der gesellschaft (17)	an der uni (35)
in unserer gesellschaft (17)	in der arbeitswelt (23)
den ganzen tag (14)	die leute die (19)
den letzten jahren (13)	dass die meisten (16)
in der heutigen (13)	in der vergangenheit (16)
der ganzen welt (10)	für die zukunft (14)
die menschen die (9)	in unserer gesellschaft (14)
in der schule (9)	man in der (14)
auf der welt (8)	bei der arbeit (13)
an der gesellschaft (7)	für viele leute (13)
auch heute noch (7)	in der heutigen (13)
auf der ganzen (7)	nach dem studium (13)
der heutigen gesellschaft (7)	die mehrheit von (12)
in den köpfen (7)	in der welt (12)
nicht so viel (7)	nach der uni (12)
die meisten menschen (6)	nicht so viel (12)
erst in den (6)	die mehrheit der (11)
in der politik (6)	in den letzten (11)

corpus. For example, the sequence *an dieser stelle* does not occur in WHiG at all and there. The higher frequencies of such devices in Falko-L1 suggest that native writers do indeed favour micro-discourse-structuring devices as opposed to the macro-discourse-structuring devices preferred by British learners of German.

Table 11 reveals that both native speakers and learners of German use a range of reference-marking sequences, of which the most frequent are place and time expressions. However, whereas native speakers seem to prefer expressions that point to shorter and more specific time periods, for example, *den ganzen tag* "the whole day" or *in den letzten jahren* "in recent years", the learners tend to use more frequently sequences expressing longer and less specific times such as *in der vergangenheit* "in the past" and *für die zukunft* "for the future". Another striking feature revealed here is that British learners frequently use a range of colloquial expressions such as *Leute* "people" as in *für viele leute* "for a lot of people" and *Uni* as in *an der uni* "at uni". In contrast, German speakers tend to use the more formal *Menschen* "human beings", as in *die meisten menschen* "most human beings".

Table 12 shows the 20 most frequently used stance expressions. As can be seen, both groups rely on a range of personal and impersonal phrases, of which the most frequent are personal stance expression containing the noun *Meinung* "opinion" as in *meiner meinung nach* "in my opinion" and the verb *sagen* "to say" as in *kann*

Table 12. The 20 most frequent stance expressions (raw frequencies)

Falko-L1	WHiG
meiner meinung nach (42)	meiner meinung nach (38)
auf jeden fall (16)	der meinung dass (35)
der meinung dass (16)	ist es nicht (26)
ich denke dass (15)	man sagen dass (26)
ist es nicht (11)	zu sagen dass (25)
man sagen dass (11)	sagen dass die (21)
kann man sagen (10)	glaube ich dass (19)
für sich selbst (9)	es ist nicht (18)
sagen dass sich (9)	und es ist (18)
bin der meinung (8)	auf jeden fall (17)
ich bin der (8)	aber es ist (16)
auf keinen fall (7)	sagen dass es (15)
der ansicht dass (7)	dass es nicht (14)
nicht mehr nur (7)	ist es klar (14)
nicht mehr so (7)	könnte man sagen (14)
so ist es (7)	es klar dass (13)
ein grosses problem (6)	ich glaube dass (12)
es ist also (6)	ein grosses problem (11)
kann man also (6)	meinung nach ist (11)
meinung nach ist (6)	deshalb ist es (10)

man sagen "can one say". There are also considerable differences. For example, the WHiG data includes more impersonal stance expressions whereby an opinion is either expressed through adjectival phrases with the existential *es ist* (as in *es ist klar* "it is clear", *es ist nicht* "it is not"), or through a phrase with the impersonal third person pronoun *man* "one". The latter contains in most instances the modal auxiliary *könnte* "could/might" as in *könnte man sagen* "one could/might say". This adds a degree of tentativeness and is often used as a hedging device to tone down utterances. In contrast, native speakers appear to use more personal and more direct stance-marking sequences as well as a number of intensifiers, such as *auf jeden fall* "by all means" and *auf keinen fall* "by no means". There are no instances of the modal auxiliary *könnte* "could/might" in stance expressions. Native speakers demonstrate a preference for the form of the verb in the indicative mode *kann* "can", which in contrast to *könnte* "could" has a sense of direct or real possibility.

5. Discussion and conclusions

Our findings reveal a number of distinctive features at the level of formulaic sequences which distinguish argumentative writing produced by advanced non-native learners of German from texts written by native speakers. Firstly, learners of German use more 3-grams than native-speakers. However, the range of these sequences is much smaller. Most of the frequently used 3-grams are variations of the two simple existential structures *es gibt* "there is/are" and *es ist* "it is". These have not only been documented in German learner language before (Maden-Weinberger 2009), but it has been suggested that the English constructions *there is/are* and *it is* in learner language are "universal and not [L1] language specific" (Maden-Weinberger 2009: 261). These 2-word combinations act as phrase-frames (O'Donnell et al. 2013), i.e. cores that are expanded by adding additional lexical items either to the left or to the right of the core. Such structures are grammatically correct. However, they rarely occur in the native corpus. This confirms observations made by Wray (1999: 223) that "for advanced learners, the major problem can lie in the production of perfectly grammatical utterances that are simply not the preferred idiomatic sequences used by native speakers". In contrast to (advanced) learners who seem to rely on the repetition of simpler constructions, native speakers use a greater range of expressions. This is in line with De Cock et al. (1998: 78), who observe that "advanced learners use prefabs, and in some cases even more prefabs than [native speakers] […] but the chunks they use (1) are not necessarily the same as those used by [native speakers], (2) are not used with the same frequency, (3) have different syntactic uses, and (4) fulfil different pragmatic functions".

When looking at the numbers of 3-grams directly copied from the topic, the data suggest that both L1 speakers and learners of German use the same proportions of such sequences, showing that a topical "safety blanket" is used in argumentative essays regardless of the language proficiency of the writer.

The functional analysis suggests that the learners in the context under study prefer to use more impersonal and cautious stance expressions than their German counterparts. Whereas Germans native speakers use *ich bin der meinung/ansicht dass* "I am of the opinion that", British learners show a preference for impersonal expressions and hedging devices *es ist klar* "it is clear", *man könnte sagen* "one could/might say". This appears contrary to the claim expressed in previous research (Chen & Baker 2010, Lorenz 1998, Hyland 1994, Hyland & Milton 1997) that non-native speakers tend to avoid cautious language. The present paper argues that the use of cautious language in academic and argumentative writing is rather cultural and not simply a matter of whether the writer is an L1 speaker or an L2 learner. Learners appear to use the acquired linguistic repertoire to construct statements that conform to the writing norm and tradition of their own culture. This aligns with the claims of contrastive rhetoric, that, "[w]hen writing in a foreign language, learners show a tendency to transfer not only the linguistic features of their native language but also its rhetorical conventions. These conventions pertain to such factors as the structure or units of texts, explicitness, information structure, politeness and intertextuality" (Leńko-Szymańska 2008: 94).

The last aspect that this paper has uncovered is that the majority of discourse-structuring formulaic sequences found in WHiG concerned the macro-structure of the essay (*in diesem aufsatz* "in this essay"), whereas Falko-L1 participants preferred micro-structuring sequences, such as *an dieser stelle* "at this point" and *die frage ob* "the question (of) whether". A study by Fandrych & Graefen (2002) on text-commenting devices in German and English academic articles has showed substantial differences in the way writers from both cultures construct texts. While German authors "put a lot of effort into making text organisation transparent by commenting on text structure" as they go along, the English writers "seem to prefer to imagine the text […] as an already finished product and give an overview of its structure" (Fandrych & Graefen 2002: 35). This corresponds with claims made by Hinds (1987) and Clyne (1987), that Anglo-Saxon writing tends to be reader-oriented, i.e. it follows the ideal of being "as clear and reader-friendly as possible, which means that the ideas have to be laid out explicitly and the text should contain a variety of markers to signal the writer's stance and to guide the reader through the text" (Leńko-Szymańska 2008: 94). Teutonic writing traditions, on the other hand, favour a reader-responsible approach, which places the main responsibility for retrieving textual meaning and authorial intention with the reader (c.f. Leńko-Szymańska 2008 for a broader contrastive study).

In a similar vein, WHiG participants tend to rely more heavily on macro-discourse-structuring formulaic sequences. This suggests that they imagine their essays holistically as a finished product. The preference of German native speakers for micro-structuring sequences in our data seems to confirm the tendency to imagine an essay as a series of arguments and as an ongoing process, through which they guide their readers. This would then suggest — unsurprisingly perhaps — that native and non-native speakers of German compose essays with two separate writing paradigms in mind. To further evaluate and potentially corroborate this finding, future research could compare discourse-structuring sequences in English L1 texts and English L2 texts produced by German learners.

The increased interest in corpus research into formulaic language has not only contributed to a better understanding of lexico-grammatical properties of English and the difficulties English learners face; it has also had a beneficial impact on the development of new learning resources (Ellis et al. 2008, Coxhead 2000, Paquot 2010, Simpson-Vlach & Ellis 2010). For German, however, there are no equivalents. This is a serious deficit given that writing, and particularly advanced forms of writing, pose major difficulties for learners of German in the context of university education (Jaworska 2009). Moreover, evidence suggests that, overwhelmingly, students of German struggle with lexico-grammatical choices appropriate for argumentative and academic text types rather than with the formal aspects of writing such as structuring essays or referencing bibliographic sources (Jaworska 2011). Unfortunately, most of the writing materials available for L2 German focus on the latter, while the former is neglected. With a better understanding of learners' language overuse and underuse, a quick-and-easy solution would be two-column lists of *do*s and *don't*s (see Krummes & Ensslin 2012) showing learners which words and phrases to avoid (e.g. overused *meiner meinung nach* "in my opinion" or *Leute* "people") and which ones to add to the text to create a better variety in lexical expressions (e.g. underused *ich denke* "I think" or *Menschen* "humans"/"people"). More corpus-informed teaching could include showing learners KWIC concordances of Falko-L1 typical word combinations and asking them to reflect on formulaic sequences. This could then be followed by asking learners to fill out gaps with formulaic sequences while providing them with enough co-text. Finally, learners should not only be made aware of general language use as evidenced by L1 corpora but also of dominant patterns of overuse, underuse and misuse of words and phrases found in learner language. Therefore it is not enough to present learners with vocabulary lists. The key is to draw their attention to culture-specific discourse practices and allow them to practise culture-specific language use. Learners need to be introduced to "the different practices" (Hyland & Tse 2007: 235) of the target discourse community. In our study, native speakers and learners have been shown to use different types of discourse-structuring devices;

British/Anglophone learners of German therefore need training in important cultural differences in writing styles in order to become aware especially of the role that micro-structuring devices play in the composition of essays in German.

Acknowledgements

This article is an outcome of the AHRC-DFG project "What's Hard in German?" (WHiG), grant reference number AH/H500081/1. We thank Ramesh Krishnamurthy (Aston University) and Professor Gerald Newton (The University of Sheffield) for their suggestions and comments on an earlier version of the manuscript.

References

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A. P. Cowie (Ed.), *Phraseology: Theory, Analysis, and Application* (pp. 101–122). Oxford, UK: Oxford University Press.

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of *make* in native and non-native student writing. *Applied Linguistics*, *22*(2), 173–194. DOI: 10.1093/applin/22.2.173

Belz, J. A. (2004). Learner corpus analysis and the development of foreign language proficiency. *System*, *32*(4), 577–591. DOI: 10.1016/j.system.2004.09.013

Biber, D., Conrad, S., & Cortes, V. (2004). 'If you look at'…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405. DOI: 10.1093/applin/25.3.371

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263–286. DOI: 10.1016/j.esp.2006.08.003

Chen, Y. H. & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, *14*(2), 30–49.

Clyne, M. (1987). Cultural differences in the organization of academic texts: English and German. *Journal of Pragmatics*, *11*(2), 211–247. DOI: 10.1016/0378-2166(87)90196-2

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly*, *33*(2), 185–209. DOI: 10.2307/3587717

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–38. DOI: 10.2307/3587951

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics*, *3*(1), 59–80. DOI: 10.1075/ijcl.3.1.04dec

De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mais & M. Hundt (Eds.), *Corpus Linguistics and Linguistic Theory: Papers from ICAME 20 1999* (pp. 51–68). Amsterdam, Netherlands: Rodopi.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English and Literatures (BELL), New Series 2*, 225–246.

De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on Computer* (pp. 67–79). Harlow, UK: Longman, 67–79.

Education, Audiovisual and Culture Executive Agency. (2008). Key Data on Teaching Languages at School in Europe. Brussels: P9 Eurydice. Retrieved from: http://eacea.ec.europa.eu/about/eurydice/documents/KDL2008_EN.pdf (last accessed September 2015).

Ellis, N. C., Frey, E., & Jalkanen, I. (2009). The psycholinguistic reality of collocation and semantic prosody (1): Lexical access. In U. Römer & R. Schulze (Eds.), *Exploring the Lexis-Grammar Interface* (pp. 89–114). Amsterdam, Netherlands: John Benjamins. DOI: 10.1075/scl.35.07ell

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*(3), 375–96.

Erman, B., & Warren, W. (2000). The idiom principle and the open-choice principle. *Text*, *20*(1), 29–62.

Fandrych, C., & Graefen, G. (2002). Text commenting devices in German and English academic articles. *Multilingua*, *21*(1), 17–43. DOI: 10.1515/mult.2002.002

Fix, U. (2008). *Texte und Textsorten: Sprachliche, Kommunikative und Kulturelle Phänomene*. Berlin, Germany: Frank & Timme.

Granger, S. (1998a). *Learner English on Computer*. Harlow, UK: Longman.

Granger, S. (1998b). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, Analysis, and Application* (pp. 145–160). Oxford, UK: Oxford University Press.

Granger, S. (2002). A birds-eye view of learner corpus research. In S. Granger, J. Hung & S. Petch-Pyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3–33). Amsterdam, Netherlands: John Benjamins. DOI: 10.1075/lllt.6.04gra

Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, *37*(2), 538–546. DOI: 10.2307/3588404

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective* (pp. 27–49). Amsterdam, Netherlands: John Benjamins. DOI: 10.1075/z.139.07gra

Groom, N. (2009). Effects of second language immersion on second language collocational development. In A. Barfield & H. Gyllstad (Eds.), *Researching Collocations in another Language* (pp. 21–33). London, UK: Palgrave.

Han, Z. H. (2011). Fossilization: A classic concern of SLA research. In S. Gass & A. Mackey (Eds.), *The Handbook of Second Language Acquisition* (pp. 476–490). New York, NY: Routledge.

Hinds, J. (1987). Reader versus writer responsibility: A new typology. In U. Connor & R.B. Kaplan (Eds.), *Writing across Languages: Analysis of L2 Text* (pp. 141–152). Reading, MA: Addison-Wesley.

Hyland, K. (1994). Hedging in academic writing and EAP textbooks. *English for Specific Purposes*, *13*(3), 239–156. DOI: 10.1016/0889-4906(94)90004-3

Hyland, K. (2008a). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4–21. DOI: 10.1016/j.esp.2007.06.001

Hyland, K. (2008b). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*(1), 41–62. DOI: 10.1111/j.1473-4192.2008.00178.x

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing*, *6*(2), 183–205. DOI: 10.1016/S1060-3743(97)90033-3

Hyland, K., & Tse, P. (2007). Is there an 'academic vocabulary'? *TESOL Quarterly*, *41*(2), 235–253.

Jaworska, S. (2009). *The German Language in British Higher Education: Problems, Challenges, Teaching and Learning Perspectives*. Wiesbaden, Germany: Harrassowitz.

Jaworska, S. (2011). Der Wissenschaftlichkeit auf der Spur: Zum Einsatz von Korpora in der Vermittlung des Deutschen als (fremder) Wissenschaftssprache. *Deutsch als Fremdsprache*, *4*, 235–244.

Juknevičienė, R. (2009). Lexical bundles in learner language: Lithuanian learners vs. native speakers. *KaLBOTYRa*, *61*(3), 61–72.

Kramsch, C. (1997). Wem gehört die deutsche Sprache? *Jahrbuch Deutsch als Fremdsprache*, *23*, 329–347.

Krummes, C., & Ensslin, A. (2012). Formulaic language and collocations in German essays: From corpus-driven data to corpus-based materials. *Language Learning Journal*, *40*(3), 110–127.

Kuiper, K., & Haggo, D. (1984). Livestock auctions, oral poetry, and ordinary language. *Language in Society*, *13*(2), 205–234. DOI: 10.1017/S0047404500010381

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners. *Language Learning*, *61*(2), 647–672. DOI: 10.1111/j.1467-9922.2010.00621.x

Leńko-Szymańska, A. (2008). Non-native or non-expert? The use of connectors in native and foreign language learners' texts. *Aile: Acquisition et Interaction en Langue Étrangère*, *27*, 91–108.

Lorenz, G. (1998). Overstatement in advanced learners' writing: Stylistic aspects of adjective intensification. In S. Granger (Ed.), *Learner English on Computer* (pp. 53–66). Harlow, UK: Longman.

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora". In M. Walter & P. Grommes (Eds.), *Fortgeschrittene Lernervarietäten* (pp. 119–140). Tübingen, Germany: Niemeyer.

Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K., & Walter, M. (2008). Das Lernerkorpus Falko. *Deutsch als Fremdsprache*, *2*, 67–73.

Maden-Weinberger, U. (2008). Modality as indicator of L2 proficiency? A corpus-based investigation into advanced German interlanguage. In M. Walter & P. Grommes (Eds.), *Fortgeschrittene Lernervarietäten* (pp. 141–164). Tübingen, Germany: Niemeyer.

Maden-Weinberger, U. (2009). *Modality in learner German: A corpus-based study investigating modal expressions in argumentative texts by British learners of German*. (Unpublished doctoral dissertation). Lancaster University, Lancaster, UK.

Möllering, M. (2004). *The Acquisition of German Modal Particles: A Corpus-based Approach*. Bern, Switzerland: Peter Lang.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam, Netherlands: John Benjamins. DOI: 10.1075/scl.14

O'Donnell, M. B., Römer, U., & Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics*, *18*(1), 83–108. DOI: 10.1075/ijcl.18.1.07odo

Paquot, M. (2010). *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London, UK: Continuum.

Raatz, U., & Klein-Braley, C. (1982). The c-test: A modification of the cloze procedure. In T. Culhane, C. Klein-Braley & D. K. Stevenson (Eds.), *Practice and Problems in Language Testing IV* (pp. 113–138). Colchester, UK: Department of Language and Linguistics, University of Essex.

Scott, M. 2008. *WordSmith Tools Version 5*. Lexical Analysis Software. Retrieved from: http://www.lexically.net/wordsmith/version5/index.html (last accessed September 2015).

Sieber, P. (1998). *Parlando in Texten: Zur Veränderung kommunikativer Grundmuster in der Schriftlichkeit*. Tübingen, Germany: Niemeyer. DOI: 10.1515/9783110940800

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology Research. *Applied Linguistics*, *31*(4), 487–512. DOI: 10.1093/applin/amp058

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford, UK: Oxford University Press.

Timmis, I. (2002). Native-speakers norm and International English: A classroom view. *ELT Journal*, *56*(3), 240–249. DOI: 10.1093/elt/56.3.240

Wend, P. (1998). *German Interlanguage: An Analysis of Beginners' German at University Level with Implications for Strategies for Teaching Foreign Languages*. Münster, Germany: Gehring.

Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching*, *32*(4), 213–231. DOI: 10.1017/S0261444800014154

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/CBO9780511519772

Authors' addresses

Sylvia Jaworska
Department of English Language and Applied Linguistics
University of Reading
Whiteknights
Reading RG6 6AW
UK

s.jaworska@reading.ac.uk

Cédric Krummes
Centre for Global Engagement
Coventry University
Priory Street
Coventry CV1 5FB
UK

cedric.krummes@coventry.ac.uk

Astrid Ensslin
School of Creative Studies and Media
Bangor University
College Road
Bangor LL57 2DG
UK

a.ensslin@bangor.ac.uk