Tracing idiomaticity in learner language: the case of BE

Przemysław Kaszubski
School of English, Adam Mickiewicz University, Poznań, Poland

1. Introduction

It is a widely known fact that language learners, especially less advanced ones, tend to rely excessively on flexible, high-frequency, 'core' vocabulary items in their foreign language use[1]. One of such commonly overused verbs is the primary verb lemma BE, whose multifarious nature must discourage many corpus researchers from devoting it time. In this paper I attempt to construct a version of the traditional tripartite scale of idiomaticity (frozen - restricted - free combinations) in order to encode various types of occurrence of lexical BE and test the extent(s) to which particular level(s) of fixedness are responsible for the reported overuse. The matter is vital, for core words are prone to forming extensions of all kinds which, contrary to the simple 'building-block' metaphor of learner lexical performance (cf. Kjellmer 1991: 124), indicate proficiency rather than non-proficiency. Before we announce that learners overuse the commonest words and possibly give them make-up work, it is useful to find out what exactly learners do with the core lemmas[2].

The premises underlying the idiomatic chart proposed here rest on both the traditional, grammatical criteria for idiomaticity (semantic opacity, lexical/syntactic fixedness, lexical/syntactic anomaly, cf. Moon 1997: 44, Hudson 1998: 8-9) as well as on corpora-inspired views of conventionality (viz. frequency and distribution, as reported in LDOCE3) and pragmatic specialisation in discourse (formulae). Since many tendencies regarding EFL vocabulary production are transfer-related, the distribution of BE's postulated idiomaticity bands will be shown in a contrastive scheme comprising both EFL learner and control non-learner and L1-based text collections. The practical goal of these examinations is to characterise quantitatively the use of BE by Polish advanced EFL learner-writers.

An underlying methodological objective of the study is to demonstrate how the needs of learner language phraseological research fail to be served by modern, robust, corpus-driven methods of text analysis.

2. Idiomatic BE: a major challenge for corpus-driven methodology

The lemma BE poses a major challenge on corpora researchers because of its versatility and extremely high frequency. In a phraseological study, one of the first tasks that needs resolving is, of course, the separation of grammatical and semantic[3] (here also called lexical) uses of BE. Semantic BE is generally to be identified not only with the existential, intransitive uses of this verb but also with its linking (copular) functions, which likewise translate lexically into other languages, a point of importance when the impact of L1 interference on EFL language production is recalled. The two basic cases of auxiliary use (cf. Quirk et al. 1985: 129-135) to be excluded from analyses of lexical BE are: 1) 'central' passives (as opposed to semi-passives and pseudo-passives in which BE functions as a copula, e.g. 'This difficulty can *be avoided* in several ways' [central passive] vs. 'Leonard *was interested in* linguistics' [semi-passive] vs. 'The building *is* already *demolished*' [pseudo-passive]; cf. Quirk et al. 1985: 167-171); and 2) the use of BE as the progressive aspect auxiliary (e.g. 'Ann *is learning*

[1] Frequency analyses of learners' language, such as Ringbom's (1998), Altenberg's (1997) or Hasselgren's (1994), clearly point this way. Resorting to safe lexical items is a frequent communication strategy not only of learners.

[2] Although some corpus linguists consider idiomatic bonds to best operate between wordforms, I follow a lemma-based approach out of conviction, after Aitchison (1994), that the lemma is the basic lexical unit of the mental lexicon (cf. Howarth's lexemic approach, 1998).

[3] Semantic uses of verbs most often correspond to the main verb function in a clause, but can also be represented by non-finite forms (infinitives and participles) and gerunds (in non-count forms, i.e. 'being' but not 'a being').

Spanish'). Especially the first group proves extremely difficult to tackle with contemporary text-processing software.

Another source of complication for disambiguation are multi-word instances of what are called 'verbs of intermediate function': neither entirely semantic nor grammatical (Quirk et al. 1985: 96-128, 136f). Two pertinent sub-classes of such verbs are: the modal idiom 'BE to <do sth>', and the more open set of semi-auxiliaries, which include 'BE able to <do sth>' , 'BE about to <do sth>', 'BE apt to <do sth>', 'BE bound to <do sth>', 'BE due to <do sth>', 'BE going to <do sth>', 'BE likely to <do sth>', 'BE meant to <do sth>', 'BE obliged to <do sth>', 'BE supposed to <do sth>', 'BE willing to <do sth>', etc. In the performed analysis two other 'verb idioms which express modal or aspectual meaning' (Quirk et al. 1985: 143) have been supplemented: 'BE inclined to <do sth>' and 'BE allowed to <do sth>' (= 'may' or 'have permission', as applied by some Polish users). The differentiation between modal idioms and semi-auxiliaries is essential insofar as the latter approximate the lexical (linking) uses of BE.

Once all the above enumerated uses of BE can be successfully identified and set aside, we can proceed to study the remaining lexical uses, which, as with any other verb, exhibit inclinations to form idiomatic (or frozen), phraseological (restricted / collocational), and open combinations with other words.

First, the FROZEN idiomatic level of BE may be postulated as consisting of those phrases in which the verb is literally 'frozen' both lexically and formally (as a particular wordform). Such uses are few and typically specialised functionally, e.g. 'that is (to say)' (used to mark repetition), 'to be sure' (epistemic modality), or 'for the time being' (time disjunct).

Phrasal/prepositional occurrences of BE can rarely be taken as integrated semantic units, but more like instances of phrasal/prepositional complementation, e.g. 'BE around' (= 'BE available'), 'BE on' (='BE working/ running/ playing'), 'BE into <sth>' (=be interested in sth). They will be associated with the RESTRICTED collocational level, discussed below. One notable exception is the frozen perfective expression 'been around' (=having had many and varied experiences), as in 'a young executive who has been around', where the meaning of 'BE around' acquires an extended metaphorical meaning.

Published sources offer little help regarding collocational habits of BE. Collocation dictionaries (cf. Benson et al 1997, Kozłowska & Dzierżanowska 1988) present very modest entries for lexical BE, or do not present them at all. This is because BE is an 'upward collocate' (Sinclair 1991) of so many words that it makes little practical sense to list all of them. A closer look at corpus data, however, proves that a good percentage of BE tokens and types *are* somehow conditioned or conventional, i.e. that they transcend the simple slot-and-filler generative paradigm which links words according to pre-selected syntactic choices.

As mentioned, lexical BE comes in two basic variants, copular (or linking) and intransitive, the former usually outnumbering the latter significantly. One important fact about copular verbs is that they require obligatory complementation, which may be of three structural types (Quirk et al. 1985: 1171-4). Two of them are simple and prototypical: 1) by an adjective phrase ('BE <adj>'; e.g. 'the menace from the plant is serious'), and 2) by a noun phrase ('BE <noun>'; e.g. 'the movies are a form of fiction'). The third kind of complementation involves the use of a (predication) adjunct, whose most frequent surface manifestation is a prepositional phrase. This type of complementation may be functionally ambiguous, since its role may be either that of an obligatory adverbial (e.g. representing the relation 'BE <place>' or 'BE <time >') or of a subject complement resembling a noun phrase or an adjective phrase (as in the pattern 'BE of <sth>': e.g. 'BE of consequence/ substance/ importance'[4] etc.; Quirk et al. 1985: 732). Quite importantly, many of the prepositional phrases functioning as subject complements of BE are multi-word units, often internally idiomatised (i.e. displaying lexical fixedness or syntactic abnormality), with corresponding adjectival synonyms, e.g.: 'BE out of breath' (cf. 'BE breathless'), 'BE of no importance' (cf. 'BE unimportant'), 'BE not at ease' (cf. 'BE not relaxed'), 'BE in love', 'BE in good condition'.

In contrast to their function as subject complements, prepositional phrases acting as obligatory adverbials seem to merely describe circumstances relating to the subject's — a person's, object's or event's — 'being' (i.e. presence or happening). They thus appear much less tied to the verb BE, which assumes a decontextualised, intransitive, existential rather than typically copular function. Such a relation is perceptible between BE and prototypical obligatory adverbials (time, space and metaphorical space) and, perhaps less strongly, also between BE and other adjunct complements (recipient, purpose, reason, accompaniment) (cf. Quirk et al. 1985: 731).

[4] 'BE of <noun>' is an interestingly productive sub-type of prepositional-phrase subject complement.

Even less transparent/ compositional (and therefore classifiable as 'restricted') seem to be the cases of complementation by: 1) means adjuncts (often conventionalised and/or lexically fixed, or else possibly replacing a passive or different predicate; e.g. 'Transport is *by ferry*', 'Entrance is *by special invitation*', 'such contracts are (= are *signed*) with people who...'); 2) stimulus adjuncts (rare, stylistically marked, and greatly restricted by the subject, which controls the preposition following BE, e.g. 'His main *interest* was *in* sport'; 3) agent adjuncts (restricted semantically to, most typically, artistic authorship, e.g. 'The book was by an unknown writer'[5]); 4) measure adjuncts (contracting a non-prototypical sense of BE, though obviously a frequent and salient one among non-beginner English learners, e.g. 'The jacket was 10 pounds'[6]).

From the above survey of the complementation patterns of lexical BE, a general rule can be inferred that the verb tends to be followed by complements which either constitute idiomatic phrases, or restrict (specialise) BE's realm of reference (by influencing its subject collocates), or which otherwise constitute simple, ad-hoc, compositional phrases (adjectival, nominal or prepositional). I would like to propose for the first two of these types to be joined into a complementation super-pattern 'BE <idiom>', which will be henceforth associated with a RESTRICTED level of collocability of lexical BE, on the grounds that: 1) copular BE, by definition, requires a complement (or adverbial), and 2) the type of complement (or adverbial) considered is itself idiomatic. Examples of restricted collocations representing the two prototypical complementation patterns ('BE <adj>' and 'BE <noun>') will include: 1) BE + adjectival idioms or collocations (predicatively unified, often substitutable by a single verb, e.g. 'BE conditional upon <sth>', 'BE worth <(doing) sth>', 'BE alive' (cf. 'live'), 'BE fraught with <sth>', 'BE sorry for <sb>' (cf. 'sympathise'); 2) predicative pseudo-passives and semi-passives (e.g. 'BE composed of <sth>', 'BE connected with <sth>', 'BE interested in <sth>', 'BE used to <(doing) sth>', 'BE situated <somewhere>'; 3) BE + adjectival/past-participial predicate + to-clause (e.g. 'BE liable to <do sth>', 'BE reluctant to <do sth>'; 4) BE + nominal idiom (e.g. 'BE a bitter pill (for <sb>) (to swallow)', 'it BE high time', 'BE the case (with <sb/sth>)'.

Following the criteria of pragmatic specialisation and frequency, another sub-category of idiomatically RESTRICTED uses should be associated with lexicalised discourse-related formulae, which in the case of BE are quite numerous. On account of the transparent, prototypical semantics of BE and absence of (strong) lexical and syntactic restrictions operating on it, formulaic uses should not, I believe, be regarded as frozen. Table I below provides a brief summary of suitable subtypes and instances of formulae:

Table I: Restricted, discourse-conditioned phrases with lexical BE

Pattern/Subtype	Example/Sub-pattern
conventional discourse formulas and linking phrases	'that/this BE why/ the reason why...' etc. (often sentence initially) 'there is every/no reason (for <sb>) to <do sth>'
'<sth> BE that...*'	'the idea/problem/thing is that...'
'<sth> BE to <do sth>**'	'his purpose/task/approach is to <do sth>'
idiomatic referential uses	'BE so/otherwise' '<sb/sth> BE one/those that/who ...'
BE + clause: other formulae***	'<sth: the question etc.> BE whether ...' '<sth> BE how <sth> <happened>' 'it/this BE because ...'
Other formulae	'<sth> BE for <sb> to <do sth>' '<sth> BE as follows/the following' etc.

* a prominent discourse prefacing formula
** a prominent explicational formula, also common in prefacing
*** this sub-type is arguably the least restricted (formulaic) of all the ones tabulated here; it has been added on account of semantic analogy to other prefacing formulas

Apart from all the restricted occurrences, lexical BE also features certain independent stylistic/rhetorical uses that are difficult to categorise within the bounds of idiomaticity. One such type of expression are cleft and pseudo-cleft sentences, where the use of BE (in bold-type in the example

[5] Many complementations of this type may be regarded as idiomatic equivalents of the long passive, e.g. 'The book was/had been written by an unknown writer'.
[6] The pattern of such expressions may be written out as 'BE <sth>' but its semantic structure is totally incongruous with the defining quality of prototypical noun complementation, captured by sentences such as 'The prize was 10 pounds'.

below) appears a result of a transformation from an underlying non-emphatic predication (underlined below) rather than a typically lexical instance comparable to cases described earlier:

> After all, **it is** marriage, the beginning of a family **that** constitutes the very basic part of every nation and society and as such it is no longer a private affair between two people.
> **All** his people ask for **is** no more war.

Another stylistically motivated feature is the use of subject-to-subject raising with an optional infinitival phrase to be, found in copulas 'SEEM (to be)', 'APPEAR (to be)', 'TURN out (to be)', 'PROVE (to be)', or in complementations of some mental verbs, especially in the passive voice, e.g. 'BE found/thought etc. (to be)' [7]. Some of these structures may be frequent enough (the most common 'SEEM (to be) <adj>/<noun>' may yield up to 100 occurrences in a 100,000-word corpus) to skew other findings for BE, depending on whether the optional infinitive is included or excluded from global counts. By an arbitrary decision, in the findings presented below, such optional occurrences of 'to be' have been added up with total scores.

Lastly, it is posited that FREE-COMBINATIONAL uses of BE should comprise all the remaining occurrences of this verb, in particular cases of non-idiomatic complementation within the two prototypical patterns: 'BE <adj>' (including -ed and -ing adjectives) and 'BE <noun>'[8]. The third major sub-category of free combinations will be made up of the de-selected instances of obligatory but semantically ancillary adverbial complementation (adjuncts): 1) 'BE <adjunct: time, space, metaphorical space>' (e.g. 'Pure fire (the stars) is in the heavens.', 'It was 10 years ago.'); and 2) 'BE <adjunct: purpose, accompaniment, measure, etc.>' (e.g. in 'BE with <sb>; BE for <sth> (=purpose); 'BE about <sth>', as when used of a book, television programme etc.)

The existential use of 'there BE' or the use of 'BE' after the anticipatory 'it' are, in the scheme proposed here, assumed as resulting from transformations of the basic theme-rheme informational model (performed, e.g., to satisfy stylistic or contextual needs) and, unless lexicalised or specialised (e.g. 'there is every reason that ...'), such forms will be treated as free-combinational, regardless of their various detailed functions (cf. Biber et al. 1999: 951-953).

The presented stratification of the potential occurrences of the verb BE demonstrates that, even if only on account of its highly diversified complementation, it is not justifiable to apply one yardstick to all instances of lexical BE present in a text. On the contrary, BE seems to feature its own model of the idiomatic cline, whose investigation may provide useful material not only for EFL researchers.

3. Automatic interfaces to corpus-bound phraseology

It goes without saying that the technological potential an average applied corpus linguist may have at his/her disposal will fall short of resolving all the delicacies necessary to describe the idiomatic distribution of BE. At the heart of the problem lies the discrepancy in the way the basic term *collocation* is understood by applied linguists and the way in which it is implemented by corpora researchers. While the traditional, applied sense of collocation associates it with co-occurrence between items forming a syntactically interpretable unit (noun phrase, verb phrase etc.), corpus-driven methods are usually focused on what is easily countable in electronically held text. However, as we shall see, statistical results based on word spans or adjacent word clusters, although useful in surveying large bodies of text, cannot fully satisfy due to both incompleteness and overgeneralisation.

---

[7] Quirk et al. (1985: 1173) report that in both British and American English certain copular verbs (APPEAR, LOOK, FEEL, SEEM, SOUND, REMAIN, STAY, BECOME, END UP, PROVE, TURN, TURN OUT, WIND UP) prefer infinitive constructions before noun phrase complements. The statistical results collated in the present study for the most frequent of the verbs, SEEM, have not confirmed this preference in native-speaker written production, instead pointing to a generally much more widespread adjectival complementation in which the highly more frequent (by 3-5 times) option is the one without 'to be'. Interestingly, this last finding showed an opposite tendency (i.e. preference for 'SEEM to be <adj>') among both advanced and intermediate Polish learners of English.

[8] One further refinement (not pursued here) within the above group might be to isolate comment-making patterns beginning with 'it' (which touch upon discourse specialisation) in opposition to clauses containing nominal subjects (cf. 'it would be irresponsible to attempt...' vs. 'Attempting ... would be irresponsible.')

Precision and recall can be improved by using POS-tagged corpora, but, unless we put in a major effort to re-edit manually, some data will still slip through, due to systematic inaccuracy of taggers. It is practically impossible to automatise the labelling of the central passives (as opposed to semi-passives and pseudo-passives). The deep-tagging program tried for this project, TOSCA-ICLE tagger (Aarts et al 1997), despite the claimed 95-6% accuracy (de Haan 1997: 218), notoriously misinterpreted 'BE used to <doing sth>' as 'BE used to <do sth>' and likewise labelled as passive each instance of 'BE related to <sth>', 'BE concerned about <sb/sth>' or 'BE satisfied with <sth>'.

Equally failing may be attempts at automatising the retrieval of significant collocations. One of the crudest ways is to extract 'recurrent word combinations' (also called 'word clusters', 'word bundles', 'word strings' etc.). Altenberg (1993, 1998) rightly showed that many of them can exhibit important, pragmatic functions, particularly in spoken discourse; however, most are not, by definition, 'idiomatic' (Biber et al. 1999: 990), and prove difficult to interpret and sub-classify as a group. Their significance is further undermined by the fact that many genuine collocations and multi-word expressions are not contiguous (Kennedy 1998: 114) and do not form fixed word strings. Clusters are certainly appealing for large-corpus research: Biber et al. devote over 30 pages to these combinations and only about 13 to all other multi-word, idiomatic expressions (Biber et al. 1999: 990-1024). However, they can uncover only very selective and very incomplete lexical associations, hidden amongst results that better indicate dominant topics (cf. 'BE allowed to adopt children' or 'with Down's syndrome BE' in Polish learner data) or stylistic mannerisms (e.g. 'it BE obvious that', 'and that BE why') than reveal collocational bonding. Many clusters signify no units at all ('it BE', 'that they BE') but cannot be stop-listed since the commonest words (e.g. prepositions) play important roles in many other, meaningful clusters.

Another approach to automatising collocation extraction is to apply co-occurrence statistics. These express arithmetically the (relative) strength of the association bond between words that tend to appear within a specified span (window) of words. Two commonly used co-occurrence formulas (with variants frequently experimented upon) are mutual information (MI) and the Z-score (McEnery & Wilson 1996: 71), the latter sometimes replaced by the more accurate t-score.

The philosophy behind MI makes it difficult to apply in studies targeting phrases with high-frequency vocabulary. MI can be applied successfully to identifying 'idiosyncratic collocations' (Oakes 1998: 90) and those which typify domain sublanguages because it privileges 'rare events' (Oakes 1998: 177). The limited helpfulness of MI in pointing at significant collocates was confirmed by the fact that even when the collocate frequency threshold was lowered to 3 (while 5 is said to be a decent statistical minimum), tests for lemmatised BE (carried out with *WordSmith Tools*) failed to produce *any* significant results within the 4:4 span. This, in view of the stratified phraseological system introduced above, is a rather questionable result. More prolific can be MI calculations performed for each attested wordform of BE. The Polish students' essay-writing corpus displayed connections between 'I' and 'sure', 'I' and 'against' and 'I' and 'afraid', mostly, however, exhibited relations with infrequent, topic-induced nouns ('monarchism', 'delusion', 'ritual', 'centuries') and adjectives ('doubtful', 'conspicuous', 'annoying'), which have less to do with genuinely significant lexicogrammatical patterning of BE.

A measure of co-occurrence which is less 'resistant' to common collocates, is the Z-score. Running *Collgen* (a sub-program of the freeware package *TACT*) on the same corpus of Polish learners' essay-writing reported associations (p-level 0.01; span 4:4[9]) between the lemma BE and the adjectivess 'able', 'likely', 'supposed' and 'afraid'. It also pointed out the habitual co-occurrence of BE and '*there'* (a most likely indicator of heavy use of the existential 'there BE' structure, indeed popular with learners), or BE and 'concerned' (indicative of learners' frequent over-reliance on the structure 'as far as <sb/sth> BE concerned').

However, a problem with rated collocate lists (even those sorted by the Z-score or t-score) is that they only indicate potentially interesting cases that require much effort and close textual analysis to verify (e.g. the association of BE and 'there' may also imply the verb phrases 'BE there'). Even access to annotated corpora is not immediately helpful, since tagging on-the-fly may go wrong and the number of tag combinations to be queried for a particular type of association is often not entirely predictable, complicating search patterns and/or prolonging computer processing time.

In short, when precision in obtaining data for a pedagogically oriented study is at stake, reliance on automatic extraction means often proves insufficient because: 1) too much 'noise' is generated in the data, which, in the case of smaller corpora, may considerably slow down analysis; 2)

[9] This is an approximation. *Collgen* actually calculates co-occurrence statistics from generated word-clusters. In this case 2-5 word clusters (including the node word) were examined.

collocations can spread beyond the typically heuristic 4:4 span, in which case they will be blocked out, while extending the span would needlessly escalate the 'noise' effect; 3) sometimes only grouping data uncovers a meaningful kind of association, whereas co-occurrence extractors work with orthographic words and easily skip over, e.g., variants of one idiomatic expression (cf. Stubbs 1998); 4) learner data (especially at lower proficiency levels) contribute to a further lowering of the statistical 'precision' and 'recall' of automated procedures, because of grammatical, stylistic, orthographic and other mistakes and errors. These, unless annotated or corrected in advance by hand, will often confuse taggers and/or skew statistics[10].

4. Contrastive Interlanguage Analysis and the applied corpus network

Learner corpora studies benefit strongly when a multi-corpus network with native and non-native reference data can be applied. Such is the framework of Contrastive Interlanguage Analysis (CIA, Granger 1996), which involves two kinds of comparisons: 1) comparison of non-native and native varieties of the same language (e.g. to identify errors, or to trace 'foreign-soundingness' in patterns of overuse and underuse); and 2) comparison of different non-native varieties of the same language (e.g. to examine if a given IL phenomenon is bound with a given L1 background or is more universal/ developmental in nature). CIA gains further diagnostic and predictive power when connected with classical CA, carried out on translation or, as in the corpus network outlined below, parallel corpora. The English corpora gathered for this project fall into five pre-established proficiency categories, the central one of which is the advanced-EFL band containing the major Polish learner corpus, IFA-PICLE. In turn, the contrastive Polish part of the network represents two proficiency levels, expert/professional and college/secondary school learner, which mirror the native English control data.

Table II: The stratification of the corpora used in the study (word token counts: hyphens within words)

non-native English					native English	
'apprentice' corpora					'expert' corpora	
1. Inter-mediate	2. Upper-intermediate	3. Advanced		4. College	5. Professional	
Polish inter-mediate EFL	Spanish (upper-) intermediate EFL	Belgian-French advanced EFL	Polish advanced EFL	British and American college learner English	British academic writing	British and American quality press
PLLC	SPAN[11]	FREN	IFA-P(ICLE)	LOCN(ARG)	MCONC[12]	LOB&BROWN
92,712 tokens	94,965 tokens	101,442 tokens	107,990 tokens	106,255 tokens	97,914 tokens	94,421 tokens

POL-STUD	'apprentice' corpus	4. College level	Polish college compositions	103,382
POL-EXP	'expert' corpus	5. Professional level	Polish academic papers + quality-press articles	101,348 tokens

The Polish advanced EFL corpus IFA-PICLE[13] belongs, alongside SPAN, FREN and LOCNARG, to the International Corpus of Learner English (ICLE) resource, which primarily samples 500-1000-word

[10] This point was worth mentioning although BE, the simplest of verbs to use and one of the first to learn in writing, poses few problems ('where'/'were' and *'ben' for 'been' were the only reported cases).

[11] The Spanish learner corpus (SPAN), although officially regarded as 'advanced' in the ICLE Project structure, had to be relegated to a lower level as it contained many more grammatical mistakes and decisively poorer vocabulary in comparison to the other advanced-level EFL data.

[12] LOCNARG is a selection of argumentative essays written by English and American secondary school and college students; the whole resource constitutes the LOCNESS corpus (=Louvain Corpus of Native English eSSays), the primary control native corpus within the ICLE family, which, arguably, is more comparable with the non-native learner data than are professionally written text samples.

argumentative essays submitted by English university students in EFL countries (Granger 1994 and 1998 ed.). Because problems were encountered finding equivalent texts (in genre and style) and equivalent sample sizes to those in the ICLE material, the expert corpora, in particular MCONC, as well as POL-EXP, occasionally include longer and/or incomplete extracts of text cut out of larger publications. No topic homogeneity could be enforced, either, but efforts were made to include, in the first place, themes typically represented in IFA-PICLE and the other ICLE learner corpora (e.g. youth and social problems: violence, drugs, TV-addiction, etc.). PLLC is an extract from the Polish part (over 500,000 tokens) of the 10-million-word *Longman Learner Corpus* (*LLC*) including short essay writings, some of which feature personal rather than argumentative discourse (hobbies, interests, plan for the future, etc.). MCONC is a collection of manually extracted, jargon-free academic English texts derived from the *MicroConcord text collection B. Academic texts* (1993). LOB&BROWN is a collection of mostly quality- and popular-press extracts retrieved from the *LOB* and *Brown* corpora (*ICAME Collection of English Language Corpora* 1991), exclusively from text Category B ('Press: Editorial') and text Category F ('Popular Lore'), including analyses on political events, popular science articles, columns and editorials on every-day life, etc., but excluding short press reports.

5. How Polish advanced EFL writers overuse BE: selective findings

Let us begin with a few procedural remarks. Due to the unmanageably high frequency of BE in each corpus, most of the demanding disambiguation tasks[14] involving non-frozen expressions (passives, semi-auxiliaries, and restricted / idiomatic phrases) had to be performed on samples of random concordance lines (500) drawn from each English corpus. When projecting the samples-based scores onto whole corpora, approximations using the standard error (0.5-2.0%) were performed, and confidence ranges established, assuming, for the easiest fit, a normal distribution and the minimum 95% confidence level. Some of the values presented below will consequently appear as (partially overlapping) continua rather than as single scores. Instead of sophisticated statistical testing, which is often dropped in applied studies of this kind (Granger 1998a), the results (comparisons of frequencies and percentages) are assessed and commented upon impressionistically.

Secondly, due to the lack of topic homogeneity in the corpora, unexpectedly skewed and possibly topic-induced frequencies had to be identified. This was done by taking standard deviation measures for each recorded expression type and group across all the seven corpora and applying a heuristically established threshold of 2 to discriminate between proportionate and skewed distributions. The latter cases were then assessed as either instances of genuine quantitative difference or, if text inspection confirmed consistent connections with a uniquely (over)represented topic, rejected from further counts.

Quantitative results obtained for the first disambiguation stage (auxiliary vs. lexical BE) showed a clear underuse of the central passives among Polish intermediate learners, possibly resulting from a more personal and casual content of their texts. The complex semi-auxiliary structure 'BE going to <do sth>' (predominantly spoken and perhaps stylistically weak, cf. Biber et al. 1999: 489) was found a characteristic of (less proficient) learner writing that contrasted deeply with native English expert data, especially its academic variety.

Overused informal expressions will reappear throughout this section, becoming a frequent feature of many EFL-based findings. Amongst semi-auxiliaries where BE functions as a linking verb, another instance of informality is a rather significant, consistent overuse of the structure 'BE able to <do sth>', noticeable especially in the performance of advanced-level native and non-native writers. Generally high statistical frequency of this 'core phrase' (over 100 occurrences in a million words, Biber et al. 1999: 517) spreads proportionally across various registers and text-types (conversation, fiction, news, academic writing), implying that the expression is very familiar to most (foreign) students of English from early stages in their education. This familiarity may be a conducive factor for the reported overuse, since the phrase is a safe option for selection in almost any language task.

Passing on to the idiomatic uses of lexical BE, three specific expectations were developed and tested: 1) negative correlation between rising proficiency and increasing frequencies of single-word (non-idiomatic) uses and/or with underrepresentation of idiomatic BE; 2) prolific presence of favourite

[13] IFA-PICLE is an extract of the PICLE corpus, containing over 230,000 words of running text (365 essays). Full information on PICLE can be found at: http://main.amu.edu.pl/~przemka.
[14] Performed with *Concord*, one of *WordSmith Tools*.

expressions ('core phrases') in EFL learner data; and 3) traceability of (at least some of) the favourite expressions to L1 (Polish).

Table III: Lexical BE: Major summary results calculated from 500-line concordance findings

95% confidence intervals	5. Professional		4. College	3. Advanced		2. Upp-Int	1. Interm
	LOB &BR	MCONC	LOCN	IFA-PICLE	FREN	SPAN	PLLC
Estimated standardised frequency per 100,000 words							
Frozen uses	>5 <76	>24 <127	>2 <75	>27 <138	>10 <98	>25 <134	>6 <93
Restricted: BE + idiom	>314 <525	>282 <504	>325 <552	>324 <568	>299 <529	>228 <443	>188 <391
Restricted: formulae	>213 <397	>290 <514	>138 <306	>280 <512	>317 <551	>173 <367	>191 <396
Cleft sentences	>47 <159	>49 <173	>54 <177	>82 <234	>36 <152	>8 <96	>0 <75
Free combinations	>1,717 <1,990	>1,778 <2,086	>2,005 <2,290	>2,346 <2,686	>2,209 <2,528	>2,574 <2,870	>3,317 <3,611
Total:	>2,552 <2,892	>2,713 <3,115	>2,775 <3,148	>3,406 <3,793	>3,176 <3,553	>3,260 <3,657	>3,968 <4,300
Estimated % of lexical BE in a corpus							
Frozen uses	>0.2% <2.8%	>0.8% <4.4%	>0.1% <2.5%	>0.8% <3.8%	>0.3% <2.9%	>0.7% <3.9%	>0. 1% <2.3%
Restricted: BE + idiom	>11.5% <19.3%	>9.7% <17.3%	>11.0% <18.6%	>9.0% <15.8%	>8.9% <15.7%	>6.6% <12.8%	>4.5% <9.5%
Restricted: formulae	>7.8% <14.6%	>10.0% <17.6%	>4.7% <10.3%	>7.8% <14.2%	>9.4% <16.4%	>5.0% <10.6%	>4.9% <9.6%
Cleft sentences	>1.7% <5.9%	>1.7% <5.9%	>1.8% <6.0%	>2.3% <6.5%	>1.1% <4.5%	>0.2% <2.8%	>0.0% <1.8%
Free combinations	>63.1% <73.1%	>61.0% <71.6%	>67.7% <77.3%	>65.2 <74.6	>65.7% <75.1%	>74.4% <83.0%	>80.2% <87.4%
Total	100%	100%	100%	100%	100%	100%	100%

The summary results presented in Table III find a good deal of agreement with the overall proficiency-based predictions expresses in the first hypothesis. Lower-proficiency students (especially PLLC) appear to use fewer collocational idioms (i.e. 'BE + idiom') than the remaining groups (corpora), and many more free combinations. In frequency terms, it is perhaps surprising to see the restricted level best represented in the two EFL advanced corpora, which is due, perhaps surprisingly, to a high share of formulae in these texts (comparable to the level characterising English academic writing MCONC). At the same time, frequencies as well as percentage data show that the two advanced-level EFL corpora and the native learner corpus share a similarly extensive predilection for the application of free combinations. LOCNARG could have approached the position of expert English corpora much closer were it not for the lower figures recorded for formulae, particularly the prefacing structures of the type '<sth: idea, purpose etc.> BE that/to...'. However, quantitative studies of formulae are a shaky matter when EFL data are involved, since learner language has been found to feature many contextually unsuitable prefabs and so their pure calculation without fathoming the context may be misleading (cf. Granger 1998b, de Cock et al. 1998).

Passing on to the frozen and restricted levels of phraseology, the data obtained point to the presence of several expressions and collocations that are favoured by learners, and to many instances probably attributable to the Polish L1. Thus, the second and third of the formulated hypotheses have also found at least some confirmation in the tests.

With respect to the idiomatic levels of BE, among the scarcely represented frozen expressions one worth noting is Polish learners' apparent overuse of the finite clausal structure 'what is more' in the function of an addition/reinforcement adverbial, e.g.:

> It is no wonder, that then some easily influenced Poles, who have not been exposed to many such films so far, might want to try living in a similar manner. *What is more*, after watching another "Rambo-like" film an average Pole may be led to thinking that committing a crime is a part of people's existence. (IFA-PICLE)

Although a legitimate idiomatic expression (*LDOCE3*: 1628), 'what is more' (often contracted to 'what's more') is an emphatic and rather spoken, though infrequent, 'polyword' (cf. Altenberg 1998: 117 or Biber et al. 1999: 1008f, 1014f). The origins of the overuse may be sought in the stylistically rhetorical (usually written) Polish transitional phrase 'co więcej', frequently employed by writers and orators to emphasise and/or extend an argument. Both native Polish corpora consulted (expert POL-EXP and learner POL-STUD) were agreed in pointing to a stable frequency of 9-10 instances per 100,000 words for this use. Although not as high as the one attested for 'what is more' in IFA-PICLE and PLLC, the value is significant enough to indicate Polish-English cognateness (or possibly transliteration since the phrase is fairly compositional) as a likely factor enhancing the detected overuse.

Interestingly, Polish users also favour an alike finite structure 'what is more <adj: important, significant etc.>' as a sentence-initial emphasising adverbial, a use more naturally rendered by single-word adverbs like 'importantly', 'significantly'. This, too, can be traced to the habitual Polish connectors such as '(a) co ważne/najważniejsze/najistotniejsze' etc., a point which returns in the discussion of discourse formulae below.

Restricted / collocational associations are much better attested and therefore more convincing. Within the super-pattern 'BE + idiom', the following findings concerning Polish EFL essay writers are worth mentioning:

- the expression 'BE full of <sth>' appears overused by intermediate learners (12 occurrences in PLLC and 10 in SPAN), and distinguishes itself also in Polish advanced learners' writing (6); a transfer trigger mechanism is likely as the expression is more informal in English (MCONC and LOB&BROWN contain only 1 instance each) than in Polish (POL-STUD: 5 occurrences), where formal persuasive discourse readily features a directly corresponding 'być pełnym <czegoś>', as in '... świat jest pełen pokus, a natura ludzka słaba' (=the world is fraught with temptations and the human nature is feeble);
- 'BE present' appears overused in IFA-PICLE (cf. Polish semi-formal 'być obecnym', as in 'Telewizja jest obecna w życiu każdego z nas', POLSTUD) but not in PLLC (possibly due to genre inconsistency), so it is impossible to fully diagnose the case;
- the semantic set 'BE connected/associated etc. with <sb/sth>', and especially the phrase 'BE connected with <sb/sth>', appears a strong Polish learner's favourite, with IFA-PICLE recording 12 and PLLC 15 occurrences (as opposed to 0-1 in all the remaining corpora); transfer influence is only partially justified since the translational equivalent (although not a cognate), 'być związanym', registers only 3-4 times in native Polish expert and learner writing alike;
- 'BE concerned with <sth>' (='deal with sth') is in prevalent use among professional English writers (8-9), but Poles (and French Belgians) prefer for 'BE concerned' to operate in the linking structure 'as far as <sth> BE concerned', which may be the reason why they do not apply it to other contexts; frequencies show that the latter phrase may be a favourite with many EFL advanced learner populations;
- 'BE slow to <do sth>' is arguably less transparent than most other phrases (cf. the Polish 'ociągać się') and may be avoided by, or perhaps is unknown to, EFL learners, while it is recorded in moderate use among native expert writers.

With respect to discourse formulae and polywords, the following instances deserve mention:

- 'that/this BE why' appears a highly typically Polish-style linker, featuring 49 occurrences in IFA-PICLE and 56 in PLLC (other corpora: FREN 23, SPAN 13, native-speaker data 1-4, including LOCNARG). Over 30% of all the occurrences appear in the short form 'that's why'. The popularity may be L1-related (POL-EXP and POL-STUD both feature dozens of sentence-initial 'Dlatego (właśnie)', as well as several 'Z tego powodu/względu' etc., which seem translational/ pragmatic equivalents). The expression is generally not very formal and unsuitably typical of Polish EFL learners' written English discourse;
- (sentence initial) 'what is more <adj: important etc.>' registers a few times in the IFA-PICLE corpus, although no (heavy) overuse has been detected; what is interesting is that native-English sources tend to employ it in clefts (e.g. 'what is more important is that...') while adverbial uses are typically covered by adverbs ('importantly', 'significantly' etc.);

- 'BE:' (e.g. 'The question was: ....') is a discernible convention in native writing (especially academic MCONC), but most EFL learners use it twice as often, which may indicate undue overuse of this simple clause structure.

To summarise this section, the cases characterising Polish learners' habits mainly concern overuse and derive from their falling back on L1-inspired options and/or on common, familiar, spoken (or universal) phraseology. Similar stylistic infelicities may also be observed in the preferred free combinational uses of BE. For instance, existential 'there BE', perhaps more characteristic of spoken English, tends to typify lower-proficiency written performance (with PLLC and SPAN recording by far the largest frequencies). Another 'spoken habit' of Polish users, both advanced and intermediate ones, is their resorting to anticipatory 'it' clauses in preference to longer structures with full nominal subjects that usually enhance textual cohesion.

6. Conclusion

Quantitative studies of learner phraseology, marked by heavy disambiguation of instances, require finer, small-corpus based comparisons rather than coarse, corpus-driven, statistical methods. Results such as those presented above would not have been possible without painful manual analysis allowing us to reach deep into corpus data. Researchers of learner corpora should strive to capture the hard-to-retrieve covert types of 'error': the overapplication and avoidance of words, expressions, structures etc. Indeed, with more advanced learners, it is those unnaturally distributed rather than incorrectly applied items that characterise 'foreign-sounding' style prominently. Regardless of the creative side of language, much of what native users say and write is influenced by conventions which, at least in statistical terms, are also expected in learners' texts and speech, particularly at the university level. The underlying philosophy is unavoidably prescriptive, in that it presupposes native 'norms' against which learners' performance can be assessed. Respectable voices advise caution against idealising learner corpus evidence, especially when it is confronted with such norms (cf. Leech 1998). However, it seems that unless we falsely hail learner corpora research as the one and only guide to native-like competence, instead of simply naming it a contributor to more successful learning, a touch of simple, educational prescriptivism should do little harm. Especially if we beware of obvious methodological pitfalls, ask demanding questions, carefully prepare and scrutinise data and avoid drawing arrogant, foregone conclusions.

The message flowing from the presented exercise is that EFL learners do tend to overapply the simplest uses of the verb BE in comparison with native writers and that the trend is happily less marked at the advanced level than at the intermediate level. Not only free combinations, however, add to the overall impression of overuse. A number of popular collocational and frozen expressions with BE, often inspired by L1 or borrowed from spoken language, also contribute strongly.

Bibliography

--- 1995 *TACT (Textual Analysis Computing Tools). Version 2.1*. Toronto, University of Toronto.

Aarts J, Barkema H, Oostdijk N 1997 *The TOSCA-ICLE tagset. Tagging manual* [accompanying the TOSCA-ICLE tagger/lemmatiser version 1.0]. Nijmegen, University of Nijmegen.

Aitchison J 1994 *Words in the mind* [2nd ed.]. Oxford - Cambridge, Mass., Blackwell.

Altenberg B 1993 Recurrent verb-complement constructions in the London-Lund Corpus. In Aarts J, de Haan P, Oostdijk N (eds), *English language corpora: design, analysis and exploitation. Papers from the 13th international conference on English language research.* Amsterdam, Rodopi, pp 227-245.

Altenberg B 1998 On the phraseology of spoken English: the evidence of recurrent word combinations. In Cowie A P (ed), *Phraseology.* Oxford, Clarendon Press, pp 101-122.

Benson M, Benson E, Ilson R 1997 *The BBI dictionary of English word combinations.* Amsterdam - Philadelphia, John Benjamins Publishing Company.

Biber D, Johansson S, Leech G, Conrad S, Finnegan E 1999 *Longman grammar of spoken and written English*. Harlow, Pearson Education Limited.

de Cock S, Granger S, Leech G, McEnery T 1998 An automated approach to the phrasicon of EFL learners. In Granger S (ed), *Learner English on computer*. London, Addison Wesley Longman, pp 67-79.

de Haan P 1997 An experiment in English learner data analysis. In Aarts J, de Mönnink I, Wekker H (eds), *Studies in English language research and teaching: in honour of Flor Aarts*. Amsterdam - Atlanta, Rodopi, pp 215-29.

Granger S (ed) 1998 *Learner English on computer*. London, Addison Wesley Longman.

Granger S 1994 The learner corpus: a revolution in applied linguistics. *English Today* 39 (10/3): 25-29.

Granger S 1996 From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In Aijmer K, Altenberg B, Johansson M (eds), *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies Lund 4-5 March 1994*. Lund, Lund University Press, pp 37-51.

Granger S 1998a The computer learner corpus: a versatile new source of data for SLA research. In Granger S (ed), *Learner English on computer*. London, Addison Wesley Longman, pp 3-18.

Granger S 1998b Prefabricated patterns in advanced EFL writing: collocations and formulae. In Cowie A P (ed), *Phraseology*. Oxford, Clarendon Press, pp 145-60.

Howarth P A 1998 The phraseology of learners' academic writing. In Cowie A P (ed), *Phraseology*. Oxford, Clarendon Press, pp 161-86.

Hudson J 1998 *Perspectives on fixedness*. Lund, Lund University Press.

Kennedy G 1998 *An introduction to corpus linguistics*. Harlow, Addison Wesley Longman.

Kjellmer G 1991 A mint of phrases. In Aijmer K, Altenberg B (eds), *English corpus linguistics: studies in honour of Jan Svartvik*. London, Longman, pp 111-127.

Kozłowska C D, Dzierżanowska H 1988 *Selected English collocations* [Revised and enlarged edition; 1st ed. 1982]. Warszawa, Państwowe Wydawnictwo Naukowe.

Leech G 1998 Preface. In Granger S (ed), *Learner English on computer*. London, Addison Wesley Longman, pp. xiv-xx.

McEnery T, Wilson A 1996 *Corpus linguistics*. Edinburgh, Edinburgh University Press.

Moon R 1997 Vocabulary connections: multi-word items in English. In Schmitt N, McCarthy M (eds), *Vocabulary: description, acquisition and pedagogy*. Cambridge, Cambridge University Press, pp 40-63.

Oakes M P 1998 *Statistics for corpus linguistics*. Edinburgh, University Press.

Quirk R, Greenbaum S, Leech G, Svartvik J 1985 *A comprehensive grammar of the English language*. London, Longman.

Ringbom H 1998 Vocabulary frequencies in advanced learner English: a cross-linguistic approach. In Granger S (ed), *Learner English on computer*. London, Addison Wesley Longman, pp 41-52.

Scott M 1996 *WordSmith: software language tools for Windows*. Oxford, Oxford University Press.

Sinclair J 1991 *Corpus, concordance, collocation*. Oxford, Oxford University Press.

Stubbs M 1998 A note on phraseological tendencies in the core vocabulary of English. *Studia Anglica Posnaniensia* XXXIII: 399-410.

Summers D (ed) 1995 *Longman dictionary of contemporary English* [3rd Edition]. Harlow, Longman Group Ltd. [*LDOCE3*]