Studia Linguistica Universitatis Iagellonicae Cracoviensis 129 (2012)
DOI 10.4467/20834624SL.12.009.0598

JUSTYNA LEŚNIEWSKA, EWA WITALISZ
Jagiellonian University, Cracow
justyna.lesniewska@uj.edu.pl, ewital@yahoo.com

NATIVE VS. NON-NATIVE ENGLISH: DATA-DRIVEN LEXICAL ANALYSIS*

Keywords: advanced EFL use, corpus analysis of learner language, lexical features of L2 writing

Abstract

This article presents a preliminary, data-driven study of a corpus of texts written by advanced Polish learners of English, which were analysed with reference to a baseline corpus of native-speaker texts. The texts included in both corpora were produced in similar circumstances (classroom setting), with the same time and word limit, and in response to the same task. We conducted a comparative lexical analysis of the two corpora, using corpus methodology (word lists, cluster analysis, concordances, keyness) to identify the most significant differences. The most important conclusion from this study is that advanced foreign language use may differ from native-speaker language use in ways which only become visible in larger samples of language, and the differences, if analysed individually, would not be regarded as errors and would go unnoticed. There is some evidence in the study that some of these differences may be attributed to cross-linguistic influence.

1. Introduction

In second language acquisition research, there is a growing body of studies dealing specifically with learners at advanced stages of proficiency. The underlying assumption behind some of these studies is that it makes sense to characterize this group

* This paper presents an investigation which is part of the research project "Creative writing and second language acquisition", Seventh Framework Programme, Marie Curie Actions, grant number 230840. Partner institutions: University of Strathclyde (UK), Jagiellonian University (Poland), Institut Superieur des Sciences Humaines de Tunis (Tunisia).

of learners in its own right, rather than simply as language learners at a specific stage on the way to becoming fully native-like. This point is often discussed with reference to the notion of "near-native proficiency" (Ringbom 1993, 2007) and also brings to mind Cook's (2002) concept of the "language user", which emphasizes the independent, legitimate status of those who use a language without achieving native-like competence. This is an important concept not only because of the rather pessimistic assumption that language acquisition is bound to reach a plateau at a certain stage, but because of the fact that, with the increasing popularity of English as a lingua franca, more and more English language users achieve high levels of competence without ever becoming fully native-like. Advanced levels of proficiency are also the focus of investigations into ultimate attainment in SLA (Birdsong 2005), and fossilization (Han 2004, Han, Odlin 2006).

Advanced foreign language use may include such subtle phenomena as the avoidance of certain items (e.g. colloquialisms), the tendency not to experiment with language, or more normative/strict judgments on grammaticality (native speakers being more lenient). However, one observation made by researchers of advanced foreign language use is of utmost relevance to this paper: namely, that deviations from native-speaker norms in advanced learners' L2 production may be very subtle, as if "hidden", and often do not take the form of explicit errors. It follows that error analysis has its limitations as far as the analysis of advanced L2 production is concerned, because the features of language use which cumulatively may be responsible for the non-native character of the language need not necessarily be identified as errors when analysed individually. Also, certain features of advanced L2 use are more likely to affect the style and register of a text rather than correctness at the syntactic/semantic level.

Another very important point is that that such characteristics can usually be noticed only in large samples of language. Advanced L2 users may produce phrases and expressions which, considered individually, are correct, in the sense that they do not violate the L2 rules of morphology, syntax, semantics, etc. However, the cumulative effect of the use of certain phrases rather than others may give the impression of non-nativeness. Therefore, the research method which seems best suited to finding such "hidden aspects" of L2 use is the analysis of corpora of texts. When corpora of learner texts are compared against some corpus-based native speaker norms, the differences which emerge can often be described in terms of the underuse or the overuse of certain structural patterns or lexical items.

One of the first studies which very clearly and convincingly illustrated this point was carried out by Granger (1998). She investigated the use of adverbs ending in *-ly* that functioned as amplifiers of adjectives in the French sub-corpus of ICLE and in a native speaker corpus. The comparison revealed a statistically very significant underuse of amplifiers in the non-native speaker corpus, both in terms of the number of types and tokens. The most striking differences occurred in the case of three particular amplifiers: *completely*, *totally* (both overused) and *highly* (underused). At the same time, "practically none of the combinations produced was felt to be unacceptable or even awkward by native speakers" (Granger 1998: 148).

There are some other corpus-based investigations into the characteristics of advanced EFL learner language, among them a study by DeCock et al. (1998), which revealed an underuse of vagueness tags by learners, studies by Lorenz (1998, 1999), in which German learners of English were found to overuse adjectival intensification, and a study by Gilquin, Paquot (2007), which investigated the phenomenon of register confusion (features of spoken language used in academic writing). Researchers have also investigated a particular lexical item or a specific construction in learner texts, for example, Belz (2004) investigated the use of the *da*-compound by learners of German, and Yeung (2009) – the use and misuse of *besides* in EFL learner texts.

While there is little doubt that corpus linguistics offers great opportunities for investigating learner language, we have encountered certain difficulties with the use of large corpora of learner language. Firstly, nowadays large corpora of texts can be collected with relative ease provided the texts are obtained in an electronic form. This is why many collections of learner texts are made up of student essays submitted in an electronic format to the instructors. Such texts, unfortunately, may not be ideal for research purposes, as they may be influenced by the wording of other texts to an unknown extent. Indeed in an early study of a corpus of learner texts, Howarth (1996: 140) observed that learner writing, especially academic writing, is "adulterated": the learner is likely to draw on a range of phrases and expressions which occur in the sources used.

Second, for researchers interested primarily in lexical and collocational aspects of L2 use, it may be difficult to obtain useful data from a corpus analysis, even with a large corpus, because the occurrence of particular lexical items is very topic-dependent, and the topics of corpus texts are often very diverse.

2. The study

2.1. Aims and methodology

In order to exploit the potential of a corpus analysis of learner language, but overcome the two difficulties mentioned above, we decided to collect a corpus of texts written by English language users of different nationalities from different L1 backgrounds which would be much more controlled in various respects that is normally expected of learner corpora. The texts would be written in very similar circumstances: in a classroom setting, with the same time and word limit, and in response to the same task, which required the use of lexis from certain semantic fields. The use of a narrative task, as suggested by Kellerman (2001: 171), seems particularly well suited to an investigation of potential cross-linguistic effects. Such a corpus, for obvious reasons, has to be small, as all the handwritten texts need to be converted into an electronic format.

The corpus described here contains texts by learners of different L1s, but in this paper we focus on an investigation of the differences between just one of the L1 groups, namely Polish learners of English, and the baseline native speaker corpus.

Our aim was to carry out a data-driven analysis, therefore we did not focus on a specific feature of the learners' interlanguage or try to formulate predictions about the nature of learner language. Instead, we used corpus methodology (word lists, cluster analysis, concordances, keyness) to identify the most significant differences between the learners and the native-speaker writers.

The narratives were elicited by means of a one-page cartoon story. In the story, a young man, Steve, borrows an expensive car from his friend George in order to impress his date, Lara. They go for a drive and admire the beautiful views of the mountains, but when they have a puncture, Steve is forced to call George for help as he does not know to deal with the problem. George arrives on a motorcycle, changes the tyre, and then leaves with Lara, who is apparently charmed by George. The subjects were free to provide their own ending to the story.

The subjects in both the Polish group and the native speaker group were comparable in terms of age and educational background – the majority being university students in their early twenties, specialising in English, modern languages or related subjects. The most important information about the corpora is presented in Table 1.

Non-native speaker corpus (henceforth NNS corpus)	**Native speaker corpus (henceforth NS corpus)**
11 588 tokens 1 393 types	7 859 tokens 1 278 types
43 texts produced by Polish advanced learners of English	32 texts by Scottish, American, Canadian, Irish students
Mean text length: 269 words	Mean text length: 246 words

Table 1. The most important characteristics of the two corpora

In order to carry out a comparative lexical analysis of the two corpora, we used the following computer programs: AntConc 3.2.1 (Anthony) and Sketch Engine (Kilgarriff). We used AntConc to derive word lists, keyword lists, clusters (4-grams), concordances, and lists of collocates, and we made use of Sketch Engine's word sketch tool, which automatically produces one-page, corpus-derived summaries of a word's grammatical and collocational behaviour. We began with a cluster analysis, generating frequency-ordered cluster lists for 4-grams (for comments on the use of 4-grams see Biber, Barbieri 2007, Cortes 2004, Hyland 2008). We then employed a modified version of the key word analysis. In the case of this measure, the term "key words" denotes words whose frequency is unusually high in comparison with a specific norm (Scott 1999). We used the measure to compare the NNS corpus against the NS corpus.

2.2. Results

Looking at the 4-gram results, we found it intriguing that certain combinations had very high scores for the NNS corpus, but were absent from the NS list of frequent

4-grams, and there were no bundles with a similar meaning there either. These noteworthy combinations were: "fell in love with" (10 occurrences), "a good impression on" (7 hits), "in love with him" (7 occurrences), "make a good impression" (6 occurrences).

Even more interesting results were yielded by the keyness analysis. The results say a lot about the differences between the two corpora, which is why we provide the full list of the top 20 results for each corpus in Table 2.

	NNS subcorpus				NS subcorpus			
1	176	26.261	that	1	44	26.708	**out**
2	47	24.805	because	2	13	23.557	says
3	53	21.971	When	3	13	23.557	spot
4	27	21.141	few	4	21	23.075	motorcycle
5	34	19.818	**girlfriend**	5	29	22.200	**off**
6	39	15.914	After	6	22	21.682	bike
7	38	15.168	came	7	57	18.217	flat
8	14	14.496	happened	8	10	18.121	**crush**
9	53	14.442	wanted	9	52	16.635	I
10	147	13.800	He	10	27	14.588	you
11	19	13.545	best	11	8	14.497	chance
12	12	12.425	Steven	12	8	14.497	Paul
13	25	12.254	But	13	8	14.497	stranded
14	57	11.237	didn	14	11	14.084	himself
15	35	11.078	**love**	15	54	13.986	at
16	61	10.987	an	16	7	12.684	drives
17	16	10.773	fell	17	7	12.684	scenic
18	16	10.773	**impression**	18	6	10.872	boots
19	105	10.394	t	19	6	10.872	wait
20	10	10.354	became	20	32	10.309	how

Table 2. Identifying lexical differences: keyness

We selected the most striking items for analysis. These items, at which we will take a closer look in the following section of the paper, are marked in Table 2 with bold type. First, we looked at the grammatical categories represented by the words on the keyness list. What seemed most notable to us was the very high position of the prepositions *out* and *off* in the NS corpus as opposed to the complete absence of these items from the NNS corpus. We also compared the 4-gram results against the keyness list and decided that the most salient items which deserved further analysis were the words "girlfriend" and "love" in the NNS corpus together with the only lexical item from the same semantic area in the NS corpus, namely "crush". Lastly, we selected "impression" because of its strong position in both the 4-gram and the keyness results.

The most striking difference between the two corpora concerns the preposition *out* (44 hits in the NS subcorpus, which means a frequency of 5.6 per 1000 words, and 16 hits in the NNS corpus, that is 1.4 uses per 1000 words. The difference lies

not only in the number of occurrences, but also in the word combinations in which *out* occurs: while some phrasal verbs occur in both texts (*find out*, *get out*, *take out*, *figure out*), there are some combinations which are only used by native speakers, most notably *blow out* (in the context of the tyre), *ask somebody out*, and *help somebody out*, as well as some other combinations.

A similar situation can be observed with regard to the preposition *off*. The NNS corpus features the following combinations with *off*:

drive off (1),
parked off the road (1),
show off (2),
take off (a jacket) (2),
go off (1).

The uses of *off* by the NNS are correct and appropriate for the context in which they occur. Still, they are definitely different from the uses of *off* in the NS group. The concordances for *off* in the NS corpus are shown in Fig. 1.

Looking at the NS use of phrasals with *off*, it seems that the story actually offers a number of obligatory occasions for their use; in other words, phrasal verbs such as *drive off*, *drop off*, or *head off* seem to be essential in this narrative. Nevertheless, the NSS managed to tell the story without them.

Another interesting lexical item revealed by the keyness analysis is the word *crush*, which appears in the NS corpus 12 times (1.5 per 1000 words). The following phrases are provided to illustrate the contexts in which *crush* appears in the NS corpus:

For a long time, Steve has had a **crush** on a very hot girl called Lara
Steve had a **crush** on a girl named Lara
Steve had a **crush** on Lara and was desperate to impress her
Steve wanted to take a girl he had a **crush** on, Lara, for a ride
Steve eventually got over his **crush**, and enrolled in a car maintenance class
Steve suddenly felt liberated from his **crush**
Steve wanted to take his **crush** Lara, for a ride
borrow George's car to impress his new **crush**, a girl called Lara
Steve had been **crushing** on Lara for the longest time

The element of the narrative which is rendered by means of phrases including the word *crush* in the NS corpus seems to be mostly rendered in the NNS corpus by the use of the word *girlfriend* with reference to Lara, as illustrated by the quotes below:

go on a date with his new **girlfriend**
wanted to spend a romantic evening with his new **girlfriend**
expected to make a good impression on his new **girlfriend**
because I would like to take my **girlfriend** Lara on a romantic excursion

thought Steve. "Job done!" said George, brushing	**off**	his hands. "Thank you!" said Lara, as Steve watch
says George, handing Steve the keys. Steve drives	**off**	and collects Lara, who is suitably attired for th
orge agrees, handing Steve the keys. Steve drives	**off**	and picks Lara up at her place. They drive to a b
high boots perhaps, voluptuous lips…), driving	**off**	to the mountains to admire the views, holding han
eorge's car. They were on their way home, to drop	**off**	Lara, when the car got a flat tire. Steve, who ne
s spending money on taking women out. Steve drove	**off**	to pick Lara up. Then they took a scenic route an
though he was having a hard time keeping his eyes	**off**	Lara. She returned his interest – she really like
tarted to notice that Lara couldn't take her eyes	**off**	George and kept saying how nice and brilliant he
looking like an action hero. He laughed as he got	**off**	his bike, shaking his head, smiling at Lara, sayi
er own eyes as the man of her dreams stopped, got	**off**	the bike, and quickly replaced the flat tire. Now
wouldn't mind being in your shoes today!". "Hands	**off**	"was Steve's riposte. After Steve had left, Georg
rge found out where they were stranded and headed	**off**	on his motorcycle. It seemed Steve had no idea ho
im for help. George said no problem and he headed	**off**	on his motorcycle to the place where they were st
took pictures at an offroad tourist site. Heading	**off**	to drop Lara **off**, they got a flat tire and Steve
ause he held her hand. Then he wanted to drop her	**off**	but on the way they got a flat tire. Steve had no
and made plans to meet the next day. They hit it	**off**	immediately, and before long they were going stea
an offroad tourist site. Heading **off** to drop Lara	**off**	, they got a flat tire and Steve had to call Georg
iew. Soon it was time to drive back and drop Lara	**off**	, but they couldn't do it – one of the car's tires
mediately takes him up on his offer and they ride	**off**	, leaving poor Steve with the task of getting the
a bike like that. George obliges, and they ride	**off**	into the night, oblivious to the fact that Steve
ut further ado he hopped onto his bike and roared	**off**	into the distance, leaving Steve and Lara to thei
that Lara thought George was marvellous and rode	**off**	into the sunset with him, leaving Steve. Steve go
to Steve. "Thanks so much" says Steve and rushes	**off**	to collect Lara. Lara comes out of her house atti
of seeing Lara again – this is his chance to show	**off**	how cool he is in front of a gorgeous girl. Indee
work changing the flat tire. Lara and Steve stood	**off**	to the side, watching. Lara was clearly impressed
t." George's generous nature kicked in and he was	**off**	in a flash on his motorbike (having first establi
otten reverie George was a lucky man. He was well-	**off**	, had a good circle of friends, was generous by na
car. He picked Lara up as arranged, and they went	**off**	to the mountains for the day. The sun was shining
arrives safely and, in no time, has the old wheel	**off**	the car and fits the new one. Lara is very gratef

Fig. 1. The concordances for *off* in the NS corpus

He was planning a date with his **girlfriend** Lara

Steve picked up his **girlfriend** Lara

borrowed a car from his best friend George to take his **girlfriend** Lara on a date

Another interesting characteristic of the NNS corpus is the high frequency of the use of the expression *fall in love*, mostly relating to the point in the story when Lara decides to ride back with George rather than stay with Steve, as illustrated by the quotes below:

she **fell in love** with his handsome friend who saved them

Lara was so impressed that she **fell in love** with George.

Of course Lara **fell in love** with George

She also appreciated his help and **fell in love** with him.

On that day Lara and George **fell in love**.

The striking disproportions in the use of the lexical items *love*, *girlfriend* and *crush* are summed up in Table 3.

	NNS subcorpus		NS corpus	
lexical item	**number of occurrences**	**frequency per 1000 words**	**number of occurrences**	**frequency per 1000 words**
crush	**0**	0	**12**	1.5
girlfriend	**34**	2.9	**3**	0.4
love	**47**	4	**7**	0.9

Table 3. The use of the lexical items *love*, *girlfriend* and *crush* in the two corpora

Another lexical item highlighted by the keyness analysis which is worth taking a closer look at is the word *impression*. It appears 16 times in the NNS corpus (1.4 per 1000 words), while in the NS corpus it occurs only once (0.1 per 1000 words).

The combinations with *impression* in the NNS corpus are mostly correct and, if considered individually, would not be considered to deviate from native-speaker norms:

He changed the tire and immediately made a good **impression** on Lara

Steve expected to make a good **impression** on his new girlfriend

He wanted to make an **impression** on Lara, the prettiest girl in the neighborhood

decided to borrow his friend's new car to make an **impression** on Lara

Lara stared at George, who made a big **impression** on her, and Steve became a bit jealous

The single occurrence of the word *impression* in the NS corpus is similar to those quoted above:

So when Steve wanted to make an **impression** on Lara, the hot girl from the office

Impression is used only once in the NS corpus. Instead the native speaker writers use the verb *impress* to render this particular element of the story. *Impress* is used more often in the NS corpus (26 hits, 3.3 per 1000 words) than in the NNS corpus (32 hits, 2.8 per 1000 words). Some examples are provided below to illustrate this point:

Steve wanted to borrow George's car to **impress** his new crush
Steve had a crush on Lara and was desperate to **impress** her
met a very attractive girl whom he wishes to **impress**
he had explained that he needed the car to **impress** a girl

The uses of *impress* in the NNS do not seem to differ from those quoted above, as illustrated by the following examples:

Steve wanted to use the car to **impress** Lara
His goal was to **impress** her
He really wanted to **impress** his girlfriend

The above analysis shows that the two items, *make an impression on sb* and *impress sb*, occur in both corpora and are used in a similar manner. However, their distribution is strikingly different. NS writers have a preference for the verbal construction, while NNS writers favour the nominal. This finding is perhaps the most important of all the observations made in this study, as it may indicate some kind of cross-linguistic influence in the form of a preference for a structure congruent with its L1 equivalent (both *impress sb* and *make an impression on sb* are rendered in Polish as *zrobić na kimś wrażenie*, literally "make on somebody an impression"). It should be stressed once again that this kind of cross-linguistic influence would go unnoticed if samples of learner language are analysed individually.

3. Conclusions

The greatest difference between the two sets of texts seems to lie in the lexical choices of the learners as opposed to those of the native speakers. These choices do not affect the accuracy of the learners' output and would not seem unusual when considered individually. Only when a corpus of texts is considered does the preference for one lexical item (e.g. *girlfriend*) over another (*crush*) become apparent. Of these lexical choices, the most striking difference appears in the case of the use of *off* and *out* as prepositions and particles in multi-word expressions. Such multi-word expressions seem to be preferred by the native speaker writers, with the non-native writers relying on other ways of conveying the same element of meaning. As the disproportion in the use of *off* and *out* in the two corpora is very significant, it seems worthy of further investigation. Another interesting topic for a more in-depth study would be the extent of cross-linguistic influence in advanced foreign language use. In the study

presented above, the preference for a structure congruent with the L1 (make an impression on sb) over a non-congruent one (impress sb) was very clearly marked.

Finally, a comment on the research method used in the study. All in all, the collection of a small but tightly controlled corpus that was subjected to an analysis employing typical corpus tools proved useful. The texts we collected enabled us to observe interesting patterns highlighting the differences between the learners and the native speakers despite the small size of the sample, because the texts dealt with exactly the same story. The data-driven approach worked well for our purpose, as it made it possible to discover patterns which had not been predicted or expected (eg. the underuse of phrasal verbs by the learners). It can be argued that such data-driven explorations of corpora of learner texts may be useful in indicating which features of learner language merit more in-depth analyses.

The most serious limitation of the study is its scope. A short narrative task like the one used here targets a limited lexical area. However, it seems that by conducting a larger number of studies like the one above, we could widen the scope and obtain a more general picture of learner language. It stands to reason that more small-scale studies targeting other specific lexical areas will give more tangible results than large corpora.

References

Biber D., Barbieri F. 2007. Lexical bundles in university spoken and written registers. – *English for Specific Purposes* 26: 263–286.

Belz J. 2004. Learner corpus analysis and the development of foreign language proficiency. – *System* 32: 577–591.

Birdsong D. 2004. Second language acquisition and ultimate attainment. – Davies A., Elder C. (eds.) *The handbook of applied linguistics*. Oxford: 82–105.

Cook V. 2002. Background to the L2 user. – Cook V. (ed.) *Portraits of the L2 user*. Clevedon: 1–28.

Cortes V. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. – *English for Specific Purposes* 23: 397–423.

DeCock S., Granger S., Leech G., McEnery T. 1998. An automated approach to the phrasicon of EFL learners. – Granger S. (ed.) *Learner English on computer*. London, New York: 67–79.

Gilquin G., Paquot M. 2007. Spoken features in learner academic writing: Identification, explanation and solution. – Davies M., Rayson P., Hunston S., Danielsson P. (eds.) *Proceedings of the corpus linguistics conference CL2007, University of Birmingham, UK, 27–30 July 2007* [http://www.corpus.bham.ac.uk/conference/proceedings.shtml].

Granger S. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. – Cowie A.P. (ed.) *Phraseology: Theory, analysis, and applications*. Oxford: 145–160.

Han Z. 2004. *Fossilization in adult second language acquisition*. Clevedon.

Han Z., Odlin T. 2006. *Studies of fossilization in second language acquisition*. Clevedon.

Hyland K. 2008. As can be seen: Lexical bundles and disciplinary variation. – *English for Specific Purposes* 2: 4–21.

Howarth P. 1996. *Phraseology in English academic writing: Some implications for language learning and dictionary making*. Tübingen.

Kellerman E. 2001. New uses for old language: cross-linguistic and cross-gestural influence in the narratives of non-native speakers. – Cenoz J., Hufeisen B., Jessner U. (eds.) *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives*. Clevedon: 170–191.

Kilgarriff A., Rychly P., Smrz P., Tugwell D. *The sketch engine*. [Software. http://www.sketchengine.co.uk/].

Anthony L. *AntConc*. [Software. http://www.antlab.sci.waseda.ac.jp/software.html].

Lorenz G. 1998. Overstatement in advanced learners' writing: Stylistic aspects of adjective intensification. – Granger S. (ed.) *Learner English on computer*. London, New York: 53–66.

Lorenz G. 1999. *Adjective intensification – learners versus native speakers: A corpus study of argumentative writing*. Amsterdam, Atlanta.

Ringbom H. 1993. Near-nativeness and the four language skills: Some concluding remarks. – Ringbom H. (ed.) *Near-native proficiency in English*. Abo: 295–306.

Ringbom H. 2007. *Cross-linguistic similarity in foreign language learning*. Clevedon.

Scott M. 1999. *Wordsmith tools*. [Software]. Oxford (UK).

Yeung L. 2009. Use and misuse of 'besides': A corpus study comparing native speakers' and learners' English. – *System* 37: 330–342.