JEZIKOSLOVLJE
14.2-3 (2013): 385-402

UDC 811.111'37=111
Original scientific article
Received on 13.01. 2013
Accepted for publication on 27.02. 2013

**Višnja Pavičić Takač[1]**
**Morana Lukač[2]**
[1]Josip Juraj Strossmayer University
Osijek
[2]Leiden University

# How word choice matters: An analysis of adjective-noun collocations in a corpus of learner essays

Foreign language learners' choice of collocations is traditionally considered to be one of the main markers of *foreignlanguageness* (Korosadowitz-Struzynska 1980: 115), hence relevant in achieving a high degree of competence in the target language. In this study we analyse the Croatian Corpus of English Learner Essays (CELE), which consists of 298 argumentative essays written as part of the state school-leaving exam. The corpus, consisting of over 74k tokens was collected in 10 different counties, and the essays were produced in 2010 and 2011. The learner's use of *adjective-noun* collocations is compared against both findings from a native speaker corpus (BNC) and a corpus of learner English (ICLE). Instead of viewing learner usage of collocations as deficient, the claims about the overuse and the underuse of statistically significant collocations are made on the basis of joint findings from the BNC and the ICLE. This approach demonstrates how native speaker data can be used for comparison, without being the norm against which the learner data should be measured, and, on the other hand, how other learner data can help distinguish between general characteristics of learner language and L1-transfer.

**Key words:** collocations; adjective-noun; non-native speakers; English as a foreign language; corpus analysis; BNC; ICLE; CELE.

## 1. Collocations and the nativelike language production

In the last decades, a substantial body of literature has been published on the non-native speakers' deficiency in the production of collocations and other types of multi-word prefabricated items (Ozaki 2011; Nesselhauf 2005; Wray 2002, Stubbs 2002; Hill 2000; Gitsaki 1999; Bahns and Eldaw 1993; Pawley and Syder 1983). The general consensus is that, regardless of the learner level, it is the deviant usage of prefabricated patterns that clearly differentiates the non-native-speaker (NNS) variety from the native-speaker (NS) language production (cf. Jafarpour et al. 2013; Hill 2000).

A number of studies have provided a more detailed insight into the specific characteristics of usage, comprehension and processing of collocations among NNSs. Despite the fact that collocation comprehension is usually not particularly difficult for the NNSs, their production proves to be a greater challenge (Ozaki 2011: 38).

Although positive correlations have been found between lexical knowledge and collocational knowledge, and overall proficiency and collocational knowledge (Gitsaki 1999; Bonk 2000), the relationships are not as straightforward as it could be assumed. Knowledge of general lexical words exceeds the knowledge of collocations among the NNSs (Bahn and Eldaw 1993: 108; Martynska 2004: 9). Additionally, although collocational knowledge generally does increase with the level of proficiency, there are relevant variations between these two variables among individual NNSs. Correlation has also been found between the perceived proficiency and the use of collocations (cf. Boers et al. 2006).

The explanation for these phenomena and for the special place collocational knowledge seems to assume in the NNSs' language use comes from idiosyncratic nature of collocations and from the the studies on language processing. The collocation elements are highly language-specific, and NNS can, due to L1 interference, "find *eat lunch* or *take lunch* a more obvious choice than *have lunch*." (Hill 2000: 51 [italics in the original]; cf. Wray 2002: 73).

Language processing constraints and differences between NSs and NNSs in storing collocations in the mental lexicon provide the most convincing answers to the puzzles of nativelike selection and nativelike fluency (Pawley and Syder 1983). NSs seem to store collocations holistically, unlike the NNSs, who store and retrieve separately each of the collocates (Granger 1998: 145–46; Pawley and Syder 1983: 218). According to Wray (2002: 209) the NS possesses "joined up knowledge" of collocations. Thus, when speaking of a big disaster, the NS would automatically use the idiomatic term *major catastrophe* in certain contexts. The NNS, on the other hand, does not possess ‘the joined up knowledge’: when encountering the collocation *major catastrophe*, the NNS breaks it down into two units without considering that the two words form a collocation. When encountering again the collocation *major catastrophe*, the NNS will have no memory of having seen the two words used together. This explanation accounts for the fact that the exposure to large amounts of collocations seems not to improve the NNSs’ collocational competence (Cowie and Howarth 1996: 92).

## 2. Defining collocations

The studies of multi-word units, including collocations, are traditionally subsumed under two approaches (cf. Granger and Paquot 2008; Nesselhauf 2005: Ch.2), the phraseological approach and the frequency-based approach.

The two approaches differ in their criteria for defining collocations. Whereas the phraseological approach (Cowie in Nesselhauf 2005: 14–17) opts for top-down linguistic criteria, the frequency-based approach (cf. Sinclair 1987; Stubbs 2002: 29) uses bottom-up statistical criteria for defining collocations.

Collocations in the traditional phraseological approach lie in the middle of Cowie’s continuum, between the variable and transparent “free combinations” (*drink tea*) on one end, and the opaque fixed idioms (*blow the gaff*) on the other (Granger and Paquot 2008: 2).

According to the frequency-based approach a collocation is, most simply put, “a frequent co-occurrence” (Stubbs 2002: 29). More elaborately, collocations are associations between two words which occur together more frequently than expected by chance (Biber and Conrad 1999: 183).

The approach taken in this study is frequency-based. More specifically, we define collocations in terms of the number of times the collocates co-occur in the observed corpora, rather than by language intuition and native-speaker judgements (for a similar definition of collocation cf. McEnery and Hardy 2012: Ch.6.2). Our choice of the analysed collocations was additionnally based on the two categories introduced in the BBI (Benson et al. 1986). Grammatical collocations consist of an open class word and a closed class word, e.g. noun-preposition (*attitude towards*) and lexical collocations consist of two open class words, e.g. verb-noun (*compose music*). The here analysed adjective-noun collocations belong to the latter group.

## 3. Previous studies on NNSs collocational competence

Depending on the applied methodology, the studies on NNSs' use of collocations can be divided into two categories: studies based on elicitation strategies and the ones based on production data (cf. Nesselhauf 2005: 4; Fan 2009: 112). The elicitation-based studies focus on the NNSs' productive skills by applying cloze tests (Bahns and Eldaw 1993; Keshavarz and Salimi 2007) and translation tasks (Bahns and Eldaw 1993).

The number of (corpus-based) studies based on production datahas recently risen (cf. Fan 2009; González Álvarez and Doval Suárez 2011; Shih 2000).

The two largest studies so far on NNSs' use of collocations were conducted by Gitsaki (1999), who used a combination of elicitation-based and production-based approaches to investigate collocational competence of 275 adolescent Greek schoolchildren, and Nesselhauf (2005), who investigated verb-noun collocations in the German ICLE sub-corpus.

The recurrent findings of studies on NNSs collocation use can be subsumed in the following points (Nasselhauf 2005: 8): NNSs generally use fewer collocations than NSs. Non-native speakers are not aware of the restrictions in collocate choice, but they are also not aware of all possible combinations. Finally, collocations pose a problem for the NNSs, one which surpasses general vocabulary problems.

## 4. The analysis of adjective + noun collocations in CELE

### 4.1. *Data collection and procedure*

The data encompassed essays written by higher intermediate (B2) English learners as part of the state school-leaving exam.[1] The learners were instructed to write an argumentative essay of 200-250 words on a given topic (e.g. advantages and drawbacks of school uniforms, positive and negative sides of international sports events) in 75 minutes.

The hand-written essays (N=298) were manually transcribed and converted in electronic files. The illegible essays (N=2) were eliminated from the corpus. For the purpose of facilitating part-of-speech (POS) tagging, misspellings were corrected, however, all other types of deviations, including deviations in word formation

[1] The Croatian state school-leaving exam (*Matura*) is designed, organised, carried out and supervised by the National Centre for External Evaluation of Education who kindly provided us with the corpus consisting of randomly selected essays.

(**visuable*, **undecent*, **inpolite*, etc.), were preserved in the corpus. The raw corpus was annotated using Wmatrix, the web-based tool (Rayson 2009), thus enabling identification of collocations via the CLAWS POS tagger. The corpus was thereupon analysed using WordSmith Tools 5.0 (Scott 2008).

The Croatian Corpus of English Learner Essays (henceforth CELE) consists of 72,598 words. The mean value of essay length is 244 words (SD=29.44).

## 4.2. *Data analysis*

### 4.2.1. *Comparing NS and NNS collocation usage*

A number of corpus studies on NNS use of collocations engage in comparing NNS against NS corpora (Fan 2009; Shirato and Stapleton 2007; Granger 1998). However, there are two considerations which need to be taken into account when applying this methodology.

First, in comparing NNS and NS language production, the NS data are often seen as the norm from which the NNS data deviate. We need to, however, be vary of attributing 'the yardstick status' to NS reference corpora. Usage in NNS data can only be viewed as deviant when compared not only against NS varieties, but also against the language variety it belongs to: the NNS variety. As Ringbom (1998: 191-192) concludes, we can apply terms 'overuse' and 'underuse' only if usage patterns deviate in the same direction both in the (reference) NS and NNS corpora.

The second consideration is related to the general limitations of corpus data. If a particular usage item does not appear in a representative NS corpus, it would be invalid to conclude that the item is never produced by NSs; the safer conclusion is that the NS usage of the analysed item is unlikely or limited (Lorenz 1999; Hargraves 2000). The corpus data are not an absolute measure of usage, they are best understood as general indicators of speaker preferences. Studies show that there is a strong correlation between what NSs perceive as natural language and token frequency: the more frequent an expression is in a NS corpus, the more naturally-sounding it is perceived to be by the NSs (cf. Smiskova et al. 2000).

### 4.2.2. *The present study*

In the present study we compare the usage of statistically significant adjective-noun collocations in the CELE against the usage of the same collocations in a NS reference corpus: the British National Corpus (BNC)[2], and in a NNS reference corpus: the International Corpus of Learner English (ICLE [available via CQPweb, cf. Hardie 2012]).[3] The list of statistically significant adjective-noun collocations in the CELE corpus is based on collocation frequency (5 ≥ hits) and association scores[4] calculated via the log likelihood test. The log likelihood relation statistic is commonly used for identifying strongly associated word pairs (Evert 2004a: 21). The log likelihood measure is an approximation to the exact p-values of the Fisher measure, and it is widely accepted as the standard for the significance of association (Evert 2004b). The cut-off point for significance in determining a collocation in this study is 99.99% (critical LL value = 15.13).

The pairs of adjective-noun collocations were chosen for the analysis due to the scarce number of studies (Balikci 2011), dealing with this particular collocation category and due to their high relative frequency in the corpus. Investigating high frequency collocations even in a corpus of a relatively small size enables us to track more general patterns of usage.

The three corpora vary considerably in their size and the text genres of which they consist. Whereas the BNC is a large representative corpus of British English, the ICLE is genre-specific, it includes, similarly to the CELE, only argumentative essays, but unlike the CELE, the ICLE essays are written by NNSs with various L1 backgrounds. The results of our analysis need to be contextualised according to these differences among the corpora. Different topics and genres of texts influence lexical choice – which accounts for the presence or absence of lexical collocations. These differences among the data were considered in the comparisons.

In the first (quantitative) part of the analysis, we compared the adjective-noun collocation strength across the three corpora, based on the collocations identified in the CELE corpus. In the second part of the analysis, we combined the qualitative

[2] The British National Corpus (BNC) is a 100 million word corpus of written (90%) and spoken (10%) data representative of a wide cross-section of British English from the later part of the 20th century. (www.natcorp.ox.ac.uk).

[3] The International Corpus of Learner English consists of argumentative essays written by higher intermediate of English from several L1 backgrounds. http://www.uclouvain.be/en-cecl-icle.html. The version of the corpus used in this research (accessed via http://cqpweb.lancs.ac.uk) consists of 2,880,826 words and 3,823 essays. The sub-corpora include data from Bulgaria, the Czech Republic, Belgium, the Netherlands, Finland, Germany, Austria, Switzerland, Italy, Poland, Russia, Spain and Sweden.

[4] For a critical review of the available association measurements in corpus linguistics see Gries (2013).

and the quantitative approach for examining the potential differences in choice and usage of collocations.

The null hypothesis of the quantitative analysis is that the statistically significant adjective-noun collocations in CELE will prove to be statistically significant in both the BNC and the ICLE corpus. In the second part of the analysis, we perform a qualitative analysis of the collocation usage in the three corpora focusing on the potential differences in the contexts of usage.

## 4.3. *Findings and discussion*

### 4.3.1. *Quantitative analysis: Comparing collocation strength*

There are 73 adjective-noun collocations identified as statistically significant in the CELE corpus. The data in Table 1 show the discrepancies between the status of the significant collocations in CELE, and their status in the ICLE and the BNC.

Table 1. Statistically significant collocations in the CELE below the statistical significance threshold in the BNC and the ICLE (N = number of hits in the corpus).

| Collocations bellow the statistical significance threshold in the BNC (p > 0.05) | | Collocations bellow the statistical significance threshold in the ICLE (p > 0.05) | | Collocations bellow the statistical significance threshold in the BNC and the ICLE (p > 0.05) | |
|---|---|---|---|---|---|
| N | % | N | % | N | % |
| 6 | 8.22% | 9 | 12.33% | 3 | 4.11% |

Contrary to the null hypothesis, there are a number of CELE collocations which are neither significant in the ICLE, nor in the BNC. Out of the statistically significant adjective-noun collocations found in the CELE, there are more of those which prove not to be statistically significant in the ICLE than in the BNC. This can be attributed to the difference in size between the ICLE corpus and the BNC: the smaller corpus produced fewer collocations. Additionally, lexical choice is strongly influenced by text topic, which is more obvious when a reference corpus is smaller and genre-specific, such as the ICLE. However, it is clear from the table that there are three collocations which scored below the more tolerant significance threshold (p=0.05) both in the ICLE and the BNC (Table 2).

Table 2. Deviant choice of adjective (usage specific for the CELE corpus).

| Collocation | Freq. CELE | No. of texts | LL CELE | Freq. BNC | LL BNC | Freq. ICLE | LL ICLE |
|---|---|---|---|---|---|---|---|
| *poor students* | 9 | 8 | 37.518* | 2 | 0.323 | 0 | n/a |
| *rich students* | 5 | 5 | 17.865* | 0 | n/a | 0 | n/a |
| *inappropriate clothes* | 6 | 5 | 39.543* | 0 | n/a | 0 | n/a |

* p ≤ 0.0001

The significant CELE collocations in Table 2 correspond to the either rarely occurring (*poor students*) or the non-existent collocations in the other two corpora. The collocation *inappropriate clothes* is an example of a word-for-word translation from L1 (hr. *neprimjerena odjeća*). These untypical collocations, however, have their synonymous counterparts in the ICLE and the CELE.

The collocations *poor* and *rich* with the word *students* are semantically vague, and substituted by other, more specific descriptive adjectives in both the ICLE (*disadvantaged*, *poorer*) and the BNC (*lower-born*, *poorest*, *under-privileged*, *needy*, *impoverished*, *disadvantaged*). In the CELE corpus, on the other hand, these synonymous collocations are almost entirely absent (exception: *under-privileged students* [N=1]).

The L1-based collocation *inappropriate clothes* corresponds to the more specific, descriptive collocations in the BNC (*outrageous*, *wrong*, *gaudy*, *tight-fitting, close-fitting clothes*).

### 4.3.2. *Combining quantitative and qualitative analysis: Differences in choice and usage*

a) Overuse of general adjectives

Two types of analysis confirm the NNSs' tendency to overuse general adjectives. The first is the key-words analysis, where the CELE wordlist was compared with the BNC wordlist (Table 3). Key-words are the ones whose frequency is unusually high in comparison with a norm, in this case the BNC (Wordsmith Help files [Scott 2008]). In the second part, we conducted an in-depth analysis of the individual collocations, by comparing the status of the CELE collocations with their synonyms in the other two corpora, and by analysing the concordance lines.

Table 3. The overused adjectives: Top 10 key adjectives in the CELE compared to the BNC.

| Adjective | % CELE | % BNC | LL |
|---|---|---|---|
| *bad* | 0.471 | 0.015 | 1688.64 |
| *different* | 0.35 | 0.048 | 571.548 |
| *good* | 0.441 | 0.082 | 558.353 |
| *important* | 0.22 | 0.039 | 291.088 |
| *big* | 0.138 | 0.025 | 177.496 |
| **unappropriate* | 0.017 | 0 | 173.363 |
| *negative* | 0.063 | 0.005 | 152.916 |
| **unpolite* | 0.014 | 0 | 144.469 |
| **propriate* | 0.014 | 0 | 137.768 |
| *strict* | 0.043 | 0.002 | 128.677 |

Three of the key adjectives in Table 3 are deviations in word formation in learner usage (**unappropriate*, **unpolite* and **propriate*). The remaining adjectives are highly frequent in both the NNS and the NS corpora, but, nevertheless, over-represented in the language of NNSs.

The data in Tables 4 and 5 are examples of the comparison of the CELE collocations with the synonymous strong collocations in the other corpora. The collocations *big problem* and *bad feelings* both contain overused high-frequency adjectives (see Table 3). Although the adjectives in the table cannot always be used interchangeably, the frequency of use in the compared corpora does illustrate tendencies of word preference. The adjectives are ordered hierarchically, according to their LL score in the BNC.

Table 4. ADJ + *problem* collocations synonymous with *big problem*.

| Collocation | BNC freq./ 10,000 | BNC LL | ICLE freq./ 10,000 | ICLE LL | CELE freq./ 10,000 | CELE LL |
|---|---|---|---|---|---|---|
| *major problem* | 0.041 | 2442.351* | 0.083 | 159.506* | 0 | n/a |
| *main problem* | 0.031 | 1736.872* | 0.097 | 144.38* | 0.689 | 35.44* |
| *real problem* | 0.029 | 1691.12* | 0.274 | 465.467* | 0 | n/a |
| *serious problem* | 0.022 | 511.349* | 0.149 | 315.22* | 0.014 | NS |
| *big problem* | 0.013 | 511.349* | 0.128 | 217.596* | 1.653 | 93.907* |

NS = not significant; * p ≤ 0.0001

Table 5. ADJ + *feelings* collocations synonymous with *bad feelings*

| Collocation | BNC freq./ 10,000 | BNC LL | ICLE freq./ 10,000 | ICLE LL | CELE freq./ 10,000 | CELE LL |
|---|---|---|---|---|---|---|
| *negative feelings* | 0.004 | 297.875** | 0.042 | 89.036** | 0.275 | NS |
| *unwanted feelings* | 0.002 | 188.184** | 0 | n/a | 0 | n/a |
| *hard feelings* | 0.003 | 125.753** | 0 | n/a | 0.275 | NS |
| *bad feelings* | 0.002 | 103.6963** | 0.01 | 10.634* | 22.177 | 1644,087** |

NS = not significant; *p≤0.01; ** p ≤ 0.0001

In both Tables 4 and 5, the significant CELE collocations were preceded by several other synonymous collocations in the BNC according to LL scores, containing more specific adjectives.

The CELE corpus is considerably smaller than the two reference corpora, therefore, it is not surprising that we find a smaller repertoire of collocations; however, from the demonstrated preferences in Tables 4 and 5, we can safely conclude that there is a tendency in the CELE towards a smaller range of possible adjective-noun collocations and towards the usage of general, instead of more specific adjectives which are preferred in the NS data and among the more advanced NNSs.

Although the majority of significant adjective-noun collocations in the CELE also prove to be statistically significant in both the BNC and the ICLE, a more in-depth analysis demonstrates that preferences in word choice also need to be investigated for achieving a better insight into the possible deviations in this particular NNS corpus.[5]

Similar findings have been reported by Shirato and Stapleton (2007), Shih (2000) and Jullian (2000) who found that learners overuse highly frequent collocations. Such collocations often express vague ideas (*good person*) where more specific meanings (*benevolent, upright, kind, tender, understanding person*) should preferably be expressed. Ringbom (1998: 193) argues that we can generally assume higher frequencies of commonly used words and lower frequencies of fairly rarely

[5] Other examples of almost exclusive usage of general adjectives, where near synonymous more specific adjectives are used more frequently by the NSs and more advanced NNSs, include *important part* and *big part* in the CELE where more specific collocations *integral, essential, large, great, major part* in the ICLE and *integral, large, essential, major, vital, significant, substantial, crucial,* and *prominent part* in the BNC are frequent synonymous alternatives, *poor countries* (exclusively) in the CELE, where *developing countries* is a preferred synonymous collocation in both the ICLE and the BNC, etc.

used words in the language of NNSs, due to their limited vocabulary. From a learning strategy perspective, Hussein (1990:128) describes this phenomenon as the 'overgeneralization strategy', namely, language learners tend to avoid the acquisition of specific terms, and opt for subsuming a number of specific meanings under few generic terms.

b) *L1 synonyms and collocation preference*

The findings from the CELE corpus show that when there are L1 synonyms for an expression in L2, the NNSs opt for the synonymous collocation, and ignore using the other possible synonyms. The examples of the L1 synonym preference are exemplified in Tables 6 and 7.

Table 6. ADJ + *time* collocations synonymous with *free time*

| Collocation | BNC freq./ 10,000 | BNC LL | ICLE freq./ 10,000 | ICLE LL | CELE freq./ 10,000 | CELE LL |
|---|---|---|---|---|---|---|
| *spare time* | 0.037 | 2629.414* | 0.267 | 855.063* | 0 | n/a |
| *leisure time* | 0.013 | 627.095* | 0.167 | 457.972* | 0 | n/a |
| *free time* | 0.016 | 274.232* | 0.444 | 789.777* | 0.826 | 49,074* |
| * p ≤ 0.0001 | | | | | | |

Table 7. Adj+*schools* collocations synonymous with *private schools.*

| Collocation | BNC freq./ 10,000 | BNC LL | ICLE freq./ 10,000 | ICLE LL | CELE freq./ 10,000 | CELE LL |
|---|---|---|---|---|---|---|
| *public schools* | 0.022 | 1158.83* | 0.0312 | 59.122* | 0 | n/a |
| *independent schools* | 0.014 | 952.40* | 0 | n/a | 0 | n/a |
| *private schools* | 0.011 | 578.53* | 0.049 | 133.357* | 0.689 | 48,137* |
| * p ≤ 0.0001 | | | | | | |

*Private schools* and *free time* are frequent collocations in the reference corpora, however, synonymous collocations used interchangeably with the two collocations both in the BNC and the ICLE do not appear in the CELE corpus.

The collocation *free time* has statistically significant synonyms *spare time* and *leisure time* in the BNC. These synonyms are also used in the ICLE corpus, however, not in the CELE. Similarly, *private schools* is a nearly synonymous collocation in the BNC to *public schools* and *independent schools*.

The collocation *public school* in reference to the more exclusive (often boarding) independent schools is limited to British English, and is not used with the same referent in neither of the NNS corpora. In the Polish section of ICLE, one author uses *non-public schools* (word-for-word translation from Polish *szkoły niepubliczne*) in reference to *private schools*, which is, in this case, proof of negative L1 transfer.

The existing translation synonyms in L1, or congruent collocations (Nesselhauf 2005: 221ff), *privatne škole* in Croatian for *private schools* and *slobodno vrijeme* for *free time*, facilitate positive transfer, however, they also limit NNS choice of collocation.

c) *Collocations and their syntagmatic relations*

Another relevant aspect of the in-depth analysis proved to be the discrepancy in the context of collocation usage among the corpora. ‘Context’ here refers to the syntagmatic relationships, or the strings of words frequently appearing to the left or the right of the collocation.

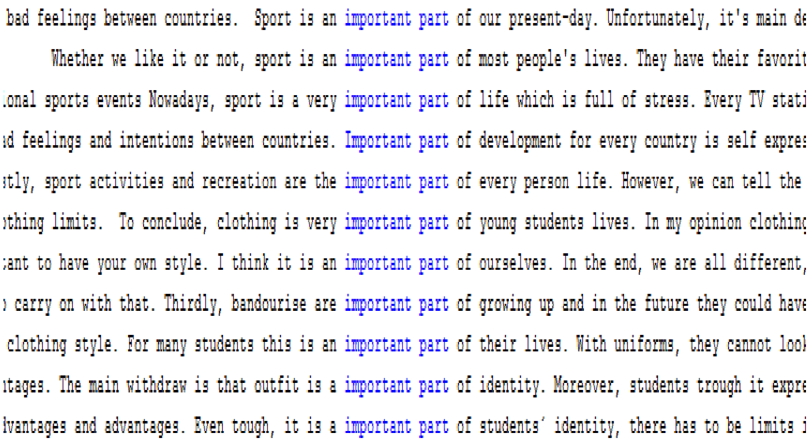
As an example of this, we here present the results of the in-depth analysis of the collocation *important part,* which is significant in all three corpora.

Table 8. 10 top collocations of *important part* in the BNC and the ICLE.

| | **BNC**[a] | | | **ICLE**[b] | | |
|---|---|---|---|---|---|---|
| No. | Word | Freq. | LL | Word | Freq. | LL |
| 1 | an | 810 | 4662.873* | an | 59 | 299.277* |
| 2 | **play** | 174 | 1407.856* | **plays** | 16 | 141.204* |
| 3 | **played** | 144 | 1297.819* | **play** | 16 | 108.923* |
| 4 | of | 621 | 766.414* | life | 21 | 74.316* |
| 5 | **plays** | 66 | 644.064* | our | 21 | 62.619* |
| 6 | in | 333 | 319.223* | of | 57 | 57.224* |
| 7 | very | 85 | 281.078* | most | 13 | 43.433* |
| 8 | most | 76 | 263.294* | **played** | 5 | 39.072* |
| 9 | is | 165 | 148.987* | very | 13 | 37.151* |
| 10 | the | 551 | 133.006* | in | 28 | 19.225* |

[a] *important part* returned 1045 matches in the BNC

[b] *important part* returned 100 matches in the ICLE



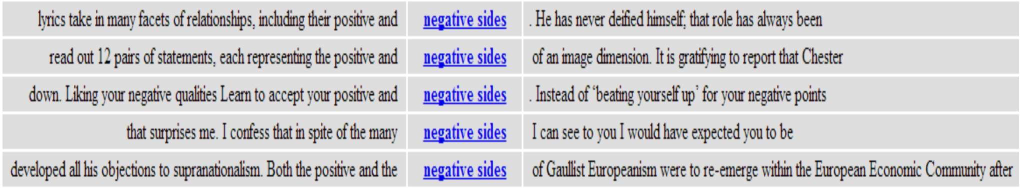
```
bad feelings between countries.  Sport is an important part of our present-day. Unfortunately, it's main de
      Whether we like it or not, sport is an important part of most people's lives. They have their favorit
.onal sports events Nowadays, sport is a very important part of life which is full of stress. Every TV stati
ıd feelings and intentions between countries. Important part of development for every country is self expres
ıtly, sport activities and recreation are the important part of every person life. However, we can tell the
ıthing limits.  To conclude, clothing is very important part of young students lives. In my opinion clothing
:ant to have your own style. I think it is an important part of ourselves. In the end, we are all different,
ı carry on with that. Thirdly, bandourise are important part of growing up and in the future they could have
clothing style. For many students this is an important part of their lives. With uniforms, they cannot look
ıtages. The main withdraw is that outfit is a important part of identity. Moreover, students trough it expre
lvantages and advantages. Even tough, it is a important part of students' identity, there has to be limits i
```

Figure 1. Concordance lines for *important part* in the CELE (Wordsmith screenshot).

Syntagmatic patterns of usage of these collocations, however, deviate in the CELE when compared against the BNC and the ICLE. Whereas *important part* appears almost exclusively preceded by the verb *to be* in the CELE (Fig. 1), the phrase *to play an important part* is highly frequent in the reference corpora (Table 8).

d) *Collocations as communication strategy*

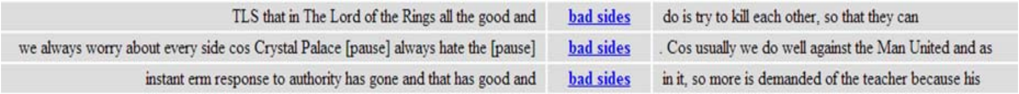
The results of the analysis further indicate that adjective-noun collocations in the CELE are occasionally used as a communication strategy[6] when a synonymous single lexical item is unknown. *Bad sides, good sides, positive sides,* and *negative sides* are significant collocations in the CELE. The use of these collocations in the BNC and in the ICLE is, however, substantially different than that in the CELE. When used in the plural form, *good sides* are usually coordinated with *bad sides*, and *positive sides* with *negative sides* (Fig. 2).



| | | |
|---|---|---|
| lyrics take in many facets of relationships, including their positive and | negative sides | . He has never deified himself; that role has always been |
| read out 12 pairs of statements, each representing the positive and | negative sides | of an image dimension. It is gratifying to report that Chester |
| down. Liking your negative qualities Learn to accept your positive and | negative sides | . Instead of 'beating yourself up' for your negative points |
| that surprises me. I confess that in spite of the many | negative sides | I can see to you I would have expected you to be |
| developed all his objections to supranationalism. Both the positive and the | negative sides | of Gaullist Europeanism were to re-emerge within the European Economic Community after |

Figure 2. Concordance lines for *negative sides* in the BNC (BNC Web screenshot).

[6] Communication strategies are employed by language users to overcome problems resulting from inedaquate target language knowledge, often to compensate for lack of lexical knowledge (cf. Tarone 1980; Dörnyei and Scott 1997; Poulisse 1993).



| | | |
|---|---|---|
| TLS that in The Lord of the Rings all the good and | bad sides | do is try to kill each other, so that they can |
| we always worry about every side cos Crystal Palace [pause] always hate the [pause] | bad sides | . Cos usually we do well against the Man United and as |
| instant erm response to authority has gone and that has good and | bad sides | in it, so more is demanded of the teacher because his |

Figure 3. Concordance lines for *bad sides* in the BNC (BNC Web screenshot).

In the CELE, however, these collocations are often found isolated, without their antonymous collocation pairs (Examples 1 and 2):

(1) *Fact is that international sports events have some* **bad sides***.*

*(2) Despite* **bad sides***, I think this is good for people.*

In the contexts where negative sides are used in Examples 1 and 2, single-word synonyms, such as *disadvantages* or *drawbacks* would be preferred in NS usage. Further support for the ‘communication strategy’ claim is provided by comparing the plots for the collocation *bad sides* and the word *disadvantages* and the plots for the collocation *good sides* with the plot for *advantages*. Only 12% of texts including the collocation *bad sides* also include the word *disadvantages*, and only 18% of texts including the collocation good sides include the word advantages.

From this we argue that collocations (and other multi-word units) are used as circumlocution when NNSs lack the knowledge of individual words (cf. Paribakth 1985; Willems 1987; Poulisse 1993).

## 5. Conclusions

The findings of this study indicate that some characteristics in the use of adjective-noun collocations are consistent across non-native corpora. Others, however, are specific for a particular group of learners which can be attributed to L1 transfer. The results further corroborate previous findings concerning the overuse of general adjectives as well as the tendency to use multi-word units as a communication strategy in situations where requisite linguistic items seem to be unavailable. We also demonstrated the usefulness of a more in-depth analysis of corpus data and ‘reading the concordance lines’ which enabled us to find differences in syntagmatic relations. In sum, our study was not based on an assumption that a NS reference corpus necessarily contains features that NNSs may aspire to, but it does serve as a general indicator of speaker preferences. Moreover, the comparison with another NNS corpus implied that NNS may diverge from a NS model without harming the communicative intent. These findings must, however, be interpreted in light of the limitations of the study which future research may attempt to overcome. For example, in order to alleviate the impact of essay topics on lexical choice the analysis may include comparisons with a corpus compiled of essays written by NSs on the same topics, or may be complemented by other methodology, such as acceptability judgement from NSs.

## References

Bahns, Jens, Moira Eldaw (1993). Should we teach EFL students collocations. S*ystem* 21.1: 101–114.

Balikci, Gözde (2011). The use of collocations by advanced learners of English: Noun-noun and adjective-noun collocations*. 1st International Conference on Foreign Language Teaching and Applied Linguistics (FLTAL '11) 5-7 May 2011, Sarajevo*. 513–520.

Benson, Morton, Evelyn Benson, Robert Ilson (1986). *The BBI Combinatory Dictionary of English.* Amsterdam – Philadelphia: John Benjamins.

Biber, Douglas, Susan Conrad (1999). Lexical bundles in conversation and academic prose. Hasselgård, Hilde, Signe Oksefjell, eds. *Out of Corpora*. Amsterdam – Atlanta, GA: Rodopi, 181–190.

Boers, Frank, June Eyckmans, Jenny Kappel, Hélène Stengers, Murielle Demecheleer (2006). Formulaic sequences and perceived oral proficiency: putting a Lexical Approach to the test. *Language Teaching Research* 10: 245-261.

Bonk, William J (2000). *Testing ESL leaners' knowledge of collocations*. http://www.eric.ed.gov/search. (20.3.2013).

Cowie, A.P., Peter Howarth (1996). Phraseological competence and written proficiency. Blue, George M., Rosamond Mitchell, eds. *Language and Education*. Clevedon: Multilingual Matters, 80–93.

Evert, Stefan (2004a). *The Statistics of Word Coocurrences: Word Pairs and Collocations*. PhD dissertation, University of Stuttgart.

Evert, Stefan (2004b). *An on-line repository of association measures*. <http://collocations.de/AM> (6.3.2013).

Fan, May (2009). An exploratory study of collocational use by ESL students - A task based approach. *System* 37: 110–123.

Gitsaki, Christina (1996). *The development of ESL collocational knowledge*. A thesis submitted for Ph.D. in the Center for Language Teaching and Research at the University of Queensland. (http://www.cltr.uq.oz.au:8000/users/christia.gitsaki/thesis/contents. html.)

González Álvarez, Elsa M., Susana M. Doval Suárez (2011). Take + noun sequences in native and learner written data. *RæL* 10: 55–68.

Granger, Sylviane (1998). Prefabricated patterns in advanced EFL writing: collocations and lexical phrases. Cowie, A.P., ed. *Phraseology: Theory, Analysis and Applica-*

*tions*. Oxford: Oxford University Press, 145–160.

Granger, Sylviane, Magali Paquot (2008). Disentangling the phraseological web. Granger, Sylviane, Fanny Meunier, eds. *Phraseology: An interdisciplinary perspective*. Amsterdam – Philadelphia: John Benjamins, 28–49.

Gries, Stefan Th. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics* 18.1: 137–165.

Hardie, Andrew (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17.2: 380–409.

Hargreaves, Peter (2000). Collocation and testing. Lewis, Michael, ed. *Teaching Collocation: Further Developments in the Lexical Approach*. London: Language Teaching Publications, 205–23.

Hill, Jimmie (2000). Revising priorities: From grammatical failure to collocational success. Lewis, Michael, ed. *Teaching Collocation: Further Developments in the Lexical Approach*. London: Language Teaching Publications, 47–69.

Hoffman, Sebastian, Hans Martin Lehmann (2000). Collocational evidence from the British National Corpus. Kirk, John, ed. *Corpora Galore: Analyses and Techniques in Describing English*. Amsterdam: Rodopi, 17–32.

Hussein, Riyad Fayez (1990). Collocations: The missing link in vocabulary acquisition amongst EFL learners. Fisiak, Jack, ed. *Papers and studies in contrastive linguistics, Volume 26. The Polish-English contrastive project*: 123-136.

Jafarpour, Ali Akbar, Mashmood Hashemian, Sepideh Alipour (2013). A corpus-based approach toward teaching collocation of synonyms. *Theory and Practice in Language Studies* 3.1: 51–60.

Jullian, Paula (2000). Creating word-meaning awareness. *ELT Journal* 54.1: 37–46.

Keshavarz, Mohammad Hossein, Hossein Salimi (2007). Collocational competence and cloze test performance: a study of Iranian EFL learners. *International Journal of Applied Linguistics* 17.1: 81–92.

Korosadowicz-Struzynska, M. (1980). Word collocations in FL vocabulary instruction. *Studia Anglica Posnaniensia* 12: 109–120.

Lewis, Michael, ed. (2000). *Teaching Collocation: Further Developments in the Lexical Approach*. London: Language Teaching Publications.

Lorenz, Gunter R. (1999). *Adjective Intensification – Learners versus Native Speakers: A Corpus Study of Argumentative Writing*. Amsterdam: Rodopi.

Martyńska, Małgorzata (2004). Do English language learners know collocations? *Investigationes Linguisticae* 11: 2–12.

McEnery, Tony, Andrew Hardy (2012). *Corpus Linguistics.* Cambridge – New York – Melbourne et al.: Cambridge University Press.

Nesselhauf, Nadja (2005). *Collocations in a Learner Corpus*. Amsterdam – Philadelphia: John Benjamins.

Ozaki, Shigeru (2011). Teaching collocations effectively with the aid of L1. *The Language Teacher* 35.3: 37-40.

Paribakht, Tahereh (1985). Strategic competence and language proficiency. *Applied Linguistics* 6.2: 132-146.

Pawley, Andrew, Frances Hodgetts Syder (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. Richards, Jack C., Richard W. Schmidt, eds. *Language and Communication*. London: Longman, 191–226.

Poullise, Nanda (1993). A theoretical account of lexical communication strategies. Schreuder, Robert, Bert Weltens, eds. *The Bilingual Lexicon*. Amsterdam – Philadelphia: John Benjamins, 157–189.

Rayson, Paul (2009). Wmatrix: a web-based corpus processing environment. Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/.

Ringbom, Håkan (1998). High-frequency verbs in the ICLE Corpus. Antoinette Renouf, ed. *Explorations in Corpus Linguistics*. Amsterdam – Atlanta, GA: Rodopi, 191–200.

Scott, Mike (2008). *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

Shih, Rebecca Hsue-Hueh (2000). Collocation Deficieny in a Learner Corpus of English: from an overuse perspective. *PACLIC 14 Proceedings*: 281-288.

Shirato, Junko, Paul Stapleton (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research* 11.4: 393–412.

Sinclair, John (1987). *Looking Up. An Account of the COBUILD Project in Lexical Computing*. London – Glasgow: Collins Cobuild.

Sinclair, John, Susan Jones, Robert Daley, Ramesh Krishnamurthy (2004). *English Collocational Studies: The OSTI Report*. London: Continuum

Smiskova, Hana, Marjolijn Verspoor, Wander Lowie (2012). Conventionalized ways of saying things (CQOSTs) and L2 development. *Dutch Journal of Applied Linguistics* 1.1: 125–142.

Sripicharn, Passapong (2010). A study of collocation and pattern of high-frequency words in a small learner corpus. *CULI's 7th International Conference. Pathways in EIL: Explorations and Innovations in Teaching and Research*. Chulalongkorn: Chulalongkorn University Language Institute, 99–115.

Stubbs, Michael (2002). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford – Malden, MA: Blackwell.

Tarone, Elaine (1980). Communication strategies, foreigner talk, and repair in interlanguage. *Language Learning* 30.2: 417–431.

Willems, Gérard (1987). Communication strategies and their significance in foreign language teaching. *System* 15: 351–364.

Wray, Alison (2002). *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

## Authors' address:

Višnja Pavičić Takač
Faculty of Humanities and Social Sciences
Jägerova 9
31000 Osijek, Croatia
vpavicic@ffos.hr

Morana Lukač
Leiden University
Leiden University Centre for Linguistics
P.N. van Eyckhof 3
2311 BV Leiden
Netherlands
m.lukac@hum.leidenuniv.nl

### O VAŽNOSTI ODABIRA RIJEČI: ANALIZA KOLOKACIJA PRIDJEV-IMENICA U KORPUSU UČENIČKIH ESEJA

Odabir i uporaba kolokacija često se ističe kao jedan od pokazatelja inojezičnosti (engl. *foreignlanguageness*, usp. Korosadowitz-Struzynska 1980), ali ujedno i visoke razine jezične sposobnosti. U ovome je istraživanju analiziran Hrvatski korpus eseja učenika engleskoga jezika (CELE) sastavljen od 198 eseja koje su napisali učenici engleskoga jezika iz deset županija 2010. i 2011. godine kao dio Državne mature iz engleskoga jezika. Korpus čini preko 74 000 pojavnica. Uspoređuje se uporaba kolokacije pridjev-imenica s nalazima iz korpusa izvornih govornika (BNC) i korpusa učenika engleskoga jezika (ICLE). Uporaba kolokacija ne karakterizira se isključivo kao manjkava (usp. Shih 2000), nego se promatra pretjerana ili rijetka uporaba statistički značajnih kolokacija u usporedbi s nalazima iz korpusa BNC i ICLE. Taj pristup pokazuje s jedne strane da se jezična uporaba izvornih govornika ne mora primijeniti u usporedbi s jezičnom proizvodnjom neizvornih govornika isključivo kao norma, a s druge strane da podaci drugih učenika mogu poslužiti u razlikovanju općih obilježja međujezika i prijenosa iz materinskoga jezika (Ringbom 1998: 191).

**Ključne riječi:** kolokacije; pridjev-imenica; neizvorni govornici; engleski kao strani jezik; analiza korpusa; BNC; ICLE; CELE.