Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 182 (2015) 433 – 440

Procedia
Social and Behavioral Sciences

4th WORLD CONFERENCE ON EDUCATIONAL TECHNOLOGY RESEARCHES, WCETR-2014

# Lexical Collocations (Verb + Noun) Across Written Academic Genres In English

Nuray Gulec[a], Bulent Arif Gulec[b]*

[a]*Ozyegin University Istanbul /Turkey*
[b]*Yildiz Technical University Istanbul /Turkey*

**Abstract**

The dominance of syntactic studies in linguistics has caused lexis and grammar to be perceived as two distinct categories. With introduction of the paradigm of cognitive linguistics, the studies in syntax have been replaced by those in lexis and concepts. Semantics has come to the fore through the studies in cognitive linguistics, and there has been a trend from syntactic studies to lexical ones. In addition to research in cognitive linguistics, construction grammar has also emphasized the continuum between lexis and grammar. With the emergence of corpus linguistics, the studies regarding the continuum between lexis and grammar have gained momentum, and thus studies of collocations have been theorized. Early studies of collocations have focused on only lexis and disregarded grammar. However, in the process the studies have also incorporated grammar as well, and this view supports the idea that each word has its own grammatical properties. Therefore, lexis and grammar should be studied on the same continuum because there is a continuum between these two categories rather than a discontinuum. Within the framework of this paradigm, this study focused on verb+noun lexical collocations across the health, physical and social sciences in the written academic genre and analyzed these lexical collocations through the frequency and chi-square analysis. The study aimed to search for commonalities and differences between the verbs with their collocations. The results showed that there were more similarities and relationship between the health and physical sciences, while the social sciences indicated a significant difference compared to the other two. The study found 165 common verbs used across the three sciences. 12 verbs among the 165 verbs were found to be candidates verb+noun lexical collocations as prototypes.

© 2015 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of Academic World Research and Education Center.

*Keywords:* Corpus linguistics, collocations, construction, prototype, written academic genre

* Bulent Arif Gulec. Tel.:. +90.0212 383 70 70
*E-mail address:* arifglc@gmail.com

1877-0428 © 2015 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of Academic World Research and Education Center.
doi:10.1016/j.sbspro.2015.04.816

## 1. Introduction

New theories of formulaic language and lexicon have been prevalent with the contributions of construction grammar and corpus linguistics. Shifting from generative grammar to formulaic language has altered perspectives pertinent to domains of language (Wray, 2002). It has been often emphasized that language, whether spoken or written, is composed of prefabricated routines and fixed expressions. One of the subcategories of formulaic language is collocations, mainly made up of grammatical and lexical units. Howarth (1998) classifies collocations as lexical and grammatical units and explicates that ‘‘lexical collocations consist of two open class words (verb + noun, adjective +noun), while collocations between one open and one closed word are grammatical’’ (p.27). The studies of lexical collocations in particular have been prolific in recent decades resulting in approaching even the term ‘collocation’ from different perspectives and distinct definitions. However, it is still one of the most controversial topics in linguistics although it is often defined as ‘a relationship between lexical items that regularly co-occur’ (Carter, 1998, p.163). Even early linguists such as Saussere (1916), Bloomfield (1933) and Firth (1951) recognized and dwelt upon the importance of collocations with similar approaches and definitions. In the same way, formulaic aspect of collocations was emphasized by other linguists as well (Hymes, 1962; Bolinger, 1976; Fillmore, 1979).

Subsequent to the diagnosis of importance of collocations, computational lexicographers (Sinclair, 1991, 1996) have empirically used collocations in their studies. These kinds of applications have led to the emergence of corpus based collocation dictionaries (Sinclair, 2004, 2005). However, it still remains a problem to determine which two words regularly co-occur in a text since one can encounter different kinds of collocations at different levels. It is quite important to make distinctions between collocations and apply the right statistical analysis while extracting collocations. Since different researchers have reached different conclusions even about the collocations of the same word, a closer look at the nature of collocations through the help of corpus linguistics is highly needed.

## 2. Research Questions

This study sought answers for the following questions:

1. What verb+ noun lexical collocations (collostructions) can be observed across academic genres?
2. How are these lexical collocations (collostructions) constructed from a constructionist grammar view?
3. Is it possible to discover prototypical lexical collocations (collostructions) according to the academic genre?

The first question intends to seek an answer to the types of verb+noun collocations across different written academic disciplines. The second question aims to analyze these collocations from constructionist and collostructionist perspective. The last question purports to find out whether prototypical lexical collocations can be extracted and elicited from the distinct academic disciplines.

## 3. Method

Table 1. Research Type and Stages of the Study

| Stages | Process | Research type |
|---|---|---|
| Stage 1 | Selection of articles from journals | Corpus-based approach |
| Stage 2 | Formation of corpora from 249 research articles from 44 journals of health, physical and social sciences. | Corpus-based approach |
| Stage 3 | Conversion of corpus into text format | Corpus-based approach |
| Stage 4 | Automatic generation of frequency lists | Descriptive analysis |
| Stage 5 | Selection of meaningful lexical collocations manually | Corpus-based approach |
| Stage 6 | Application of statistical analyses across and within corpora (Fisher`s exact test) | Quantitative corpus analysis |
| Stage 7 | Checking prototypical lexical collocations | Descriptive and Interpretative |
| Stage 8 | Analysis of prototypical lexical collocations through construction grammar | Interpretative |

### *3.1 Data gathering procedure*

The database of this study was formed from 249 research articles from 44 journals of health, physical and social sciences.

### *3.2 Written Academic Corpora*

The corpora for this study were retrieved from internationally recognized, electronic journals research articles (RA). A corpus of 249 research articles (116 for health: 84 for physical, and 49 for social sciences) included 1,217.197 words. Each science type was planned to have the similar number of words. Therefore, the number of articles varied but the number of the words for each science remained similar (see Table 2). Recent articles published between 2009 and 2011 were chosen.

Only professional texts were chosen from the journals of three mainstream sciences to gain an insight into the analysis of across and within disciplines.

Table 2 The Overall Data of the Texts

| Science type | Number of disciplines | Number of research articles | Years | Total words |
|---|---|---|---|---|
| Health science | 20 | 116 | 2009-2011 | 405,753 |
| Physical science | 14 | 84 | 2009-2011 | 405,751 |
| Social science | 10 | 49 | 2009-2011 | 405,693 |
| Totals | 44 | 249 | 2009-2011 | 1,217.197 |

Table 2 indicates that the number of the words in each genre was rendered almost equal so that more reliable results could be obtained between and across the genres. The number of the disciplines varied because each discipline has a different number of pages and words. However, the number of the words remained similar. The disciplines for each genre are shown in Table 3.

Table 3 Disciplines Chosen for the Corpora in Three Distinct Sciences

| Health science | Physical science | Social science |
|---|---|---|
| Anatomy | Agriculture-Plant sciences | Literature |
| Anesthesiology | Astronomy | Anthropology |
| Bacteriology | Bioengineering | Education |
| Brain-Neuroscience | Botany | Gay and lesbian studies |
| Cardiology | Chemistry | Law |
| Cell Biology | Chemical and Materials | Philosophy |
| Dentistry | engineering | Political Science |
| Dermatology | Civil Engineering | Psychology |
| Endocrinology | Environmental Sciences | Recreation and Sports |
| Gastroenterology | Geology | Sociology |
| Genetics | Marine Science | |
| Geriatrics | Mechanical Engineering | |
| Immunology | Meteorology and Climatology | |
| Internal medicine | Physical Geography | |
| Nephrology | Physics | |
| Ophthalmology | | |
| Pediatric | | |
| Physiology | | |
| Psychiatry | | |
| Radiology | | |

These texts were transformed into text format in order to create an electronic corpus of 1,217.197 words. Lexical collocations, specifically verbs, were extracted from the corpus. Since the aim of this study was to analyze

verb+noun lexical collocations, other word classes were excluded. The classification of the collocations was done in accordance with the operational definition.

### *3.3 Software Programs*

Since each type of software has distinct properties, two different types of software were used in order to reach reliable results. Some software programs are available on internet and are easily accessible and easy to use at a basic level. The first software used in this study was concordance that provided the basic results (Watt, 2012). This software does not carry out detailed inferential statistics but offers basic descriptive statistics. Counting words, making word lists and word frequency lists, full concordances, choosing pick lists, using multiple input files are among the functions of this software. The second software utilized was Antconc that offers a better service because Antconc provides multi-layered results composed of clusters, concordance plot and basic statistical measurement. The basic statistical tools in Antconc are log-likelihood, average value and clustering. Although it does not present a detailed statistical measurement, it was used for the basic statistical results. The third software was Wordsmith, a relatively sophisticated and integrated corpus software program used for text processing and extracting verb+noun lexical collocations descriptively and inferentially (Scott, 2010). A few steps should be followed in order to reach expected results. In this sense, Wordsmith offers to generate word lists according to its alphabetical and frequency order, concordance, to find collocations and show frequencies altogether with statistical tools. It can also compare different texts by showing their statistical significance level. Wordsmith is highly developed software compared to the first two. It basically contains three modules composed of Concord, Keyword and Wordlist. T-score, chi-square score, Z-score and mutual information analysis can be performed through the Wordsmith software program. This software offers several important services such as generating concordances, listing occurrences and co-occurrences of the key words in a given text, comparing words and carrying out basic statistical analysis.

## 4. Results

The overall descriptive results of the key words used in the texts were given, and a summary statistics of the texts themselves was presented. Figures (4.1.), (4.2.) and (4.3.) present the summary statistics of the three disciplines.

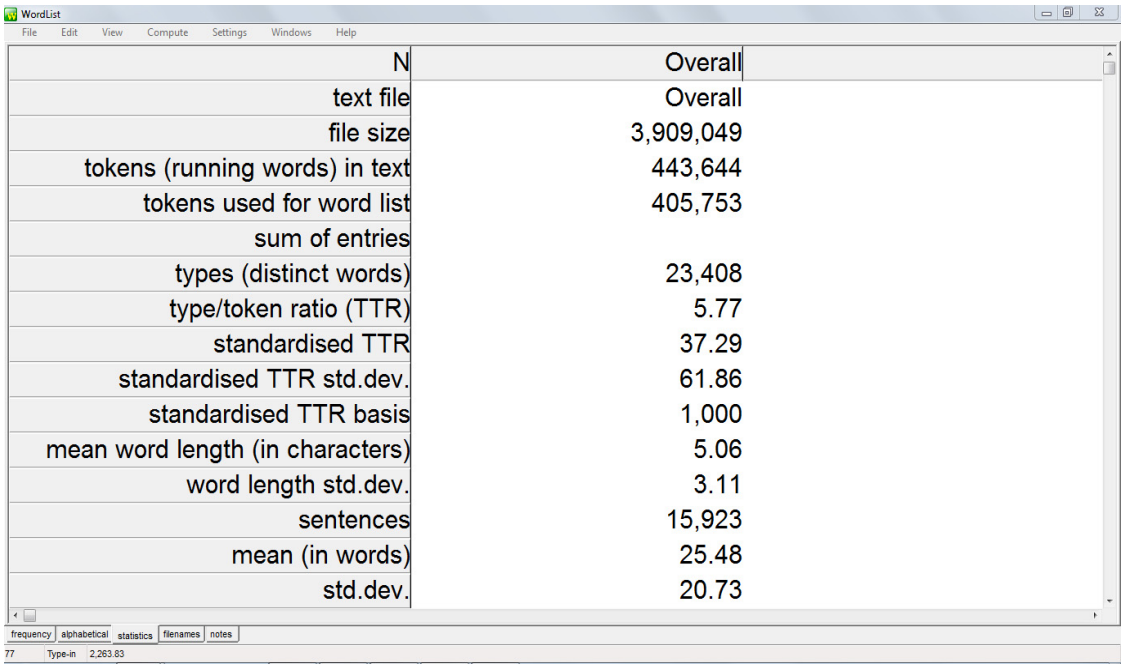



Figure1. Summary statistics of the health science texts

Table 4. Descriptive Statistics of Verbs According to the Tokens

| Academic genre | Total words | Verbs with collocates | % |
|---|---|---|---|
| Health science | 405,753 | 8740 | 2.15 |
| Physical science | 405,751 | 7298 | 1.79 |
| Social science | 405,693 | 12206 | 3.00 |
| Total | 1,217.197 | 28244 | 2.32 |

The percentage of the verbs in Table 4 shows a similar variation. The percentage of the verbs in social science is the highest (3.00%), while physical science forms the lowest percentage (1.79%). Health science accounts for only 2.15% of the verbs. Table 5 exhibits the ratio of verbs considering the types.

Table 5. The Overall Statistical Results of Verbs According to the Types

| Academic genre | Total words | Collocational verbs | % |
|---|---|---|---|
| Health science | 23.408 | 724 | 3.09 |
| Physical science | 26.717 | 556 | 2.08 |
| Social science | 23.522 | 920 | 3.91 |
| Total | 73.647 | 2190 | 2.98 |

It can clearly be seen from Table 5 that the collocational verbs in the social sciences account for the highest percentage (3.91%), whereas the verbs in the physical sciences constitute the lowest percentage (2.08%). The percentage of the verbs in the health sciences is only 3.09%. The total percentage of the verbs in terms of types is 2.98%.

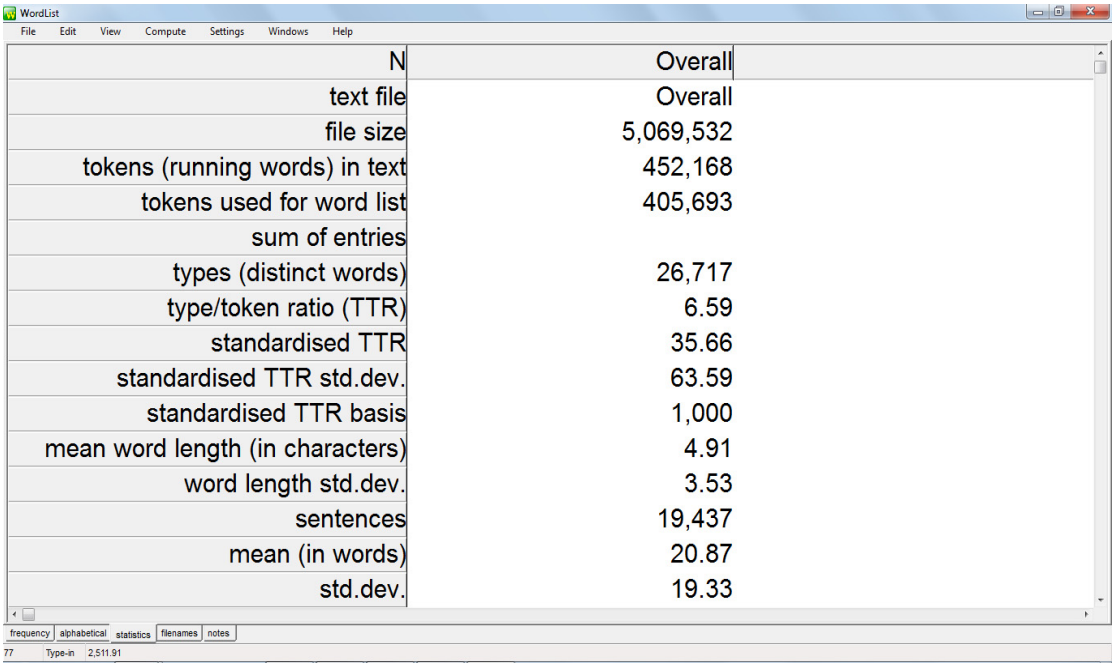
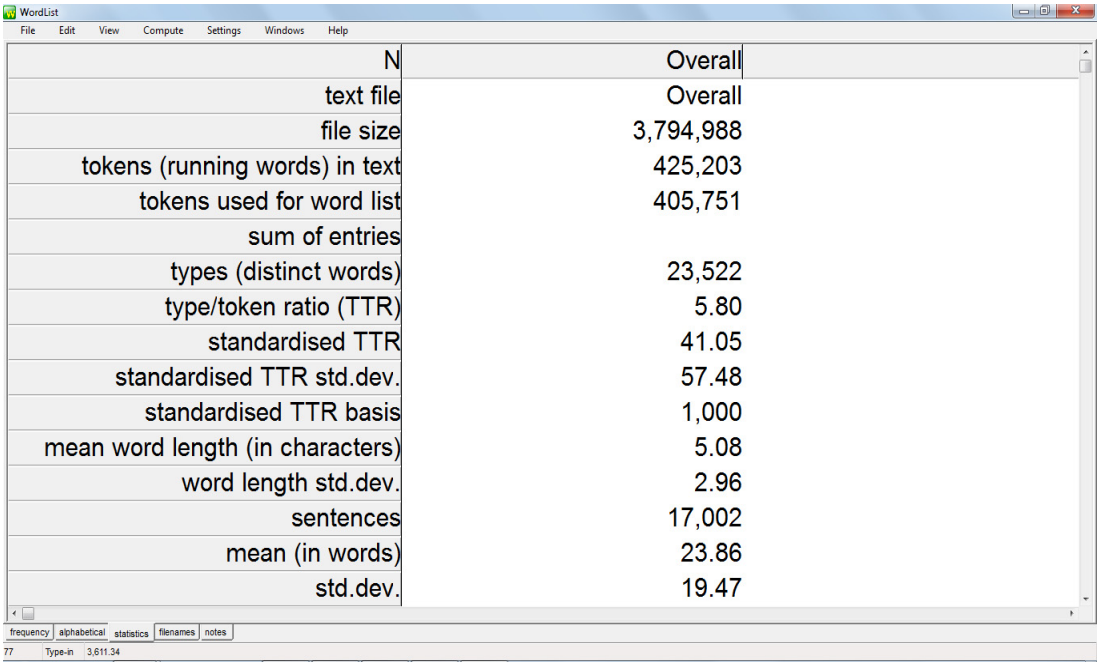



Figure 2 Summary statistics of the physical science texts




| N | Overall |
|---|---|
| text file | Overall |
| file size | 3,794,988 |
| tokens (running words) in text | 425,203 |
| tokens used for word list | 405,751 |
| sum of entries | |
| types (distinct words) | 23,522 |
| type/token ratio (TTR) | 5.80 |
| standardised TTR | 41.05 |
| standardised TTR std.dev. | 57.48 |
| standardised TTR basis | 1,000 |
| mean word length (in characters) | 5.08 |
| word length std.dev. | 2.96 |
| sentences | 17,002 |
| mean (in words) | 23.86 |
| std.dev. | 19.47 |

Figure 3.Summary statistics of the social sciences texts

## 5. Conclusion

Understanding the nature of lexical collocations in its wider sense in corpus and applied linguistics has given rise to various research questions. Some researchers even theorized the collocations in linguistics, psychology and applied linguistics. Since the foundation of collocations has not been thoroughly carried out, it is important to pose new questions, and discuss theories and ideas. This study aimed to find answers to the following research questions:

Research Question 1: What verb+noun lexical collocations (collostructions) can be observed across academic genres?: The study showed that similar verbs were used across the three academic genres: health, physical and social. However, these verbs showed some variation in terms of the collocates they attracted. Collocates in the social sciences showed more variation compared to those in the health and physical sciences. The number of verbs taking collocates was more limited in the health and physical sciences.

Research Question 2: How are these lexical collocations (collostructions) constructed from a constructionist grammar view?: The study did not intend to deal with collocations in a traditional sense. Rather, it included a question based on recent research discussions and findings. The results of this study showed that the verbs in written academic genres tended to occur with constructions besides only simply co-occurring words. Almost each verb was seen to have its collostructional properties. It was found that there were no pure verb+noun collocations in their pure and naïve form. The most frequently used verbs with their collostructions showed a similar result in several studies as well (Thompson and Ye, 1991; Hyland, 1999, 2000; Hyland and Tse, 2005). In the three academic genres, one of the strongest collostruct was found to be that-clause collostruct. This finding is important in that Goldberg (2006) stresses the importance of the frequency and entrenchment of a specific construction.

Research Question 3: Is it possible to discover prototypical lexical collocations (collostructions) according to the academic genre?: This last question intended to find out whether the data could produce some prototypicalities similar to those in linguistics and psychology. The results of the study showed that prototypes existed in the social context of written academic prose. In general 165 common possible prototypical verbs were detected, although statistically there seemed no significant relationship between the genres. Out of these 165 common verbs, 12 most frequent and most common verbs across the three genres were seen to have prototypical features at high frequencies.

As the degree of the frequency decreased, the variation of the verbs increased. This result is also supported by Hyland and Tse (2007) stressing that only 8% and 10% of the words show similar frequencies across different genres, and in terms of technical vocabulary, only 5% of the running words indicate similarities implying that genres show 'discursive variability'(p.251). It is not surprising that only a small percentage of the data show similarities because each sub-discipline produces different combinations. Therefore, Hyland and Tse (2007) approach academic vocabulary list with caution by insistently stating that these kinds of results may refer to the misrepresentation of academic literacy. Psychological explanation of conceptual combinations and linguistics explanation of collocations have shown that it is a thorny issue to find prototypes at the level of collocations (Murphy, 2004; Hyland and Tse, 2007). Hyland and Tse (2007) in their study concludes that it may be pedagogically misleading for learners to direct them to 'overarching, universally appropriate teaching items' (p.251).

**Implications and Recommendations for Language Learning and Teaching**

This study has revealed that similar verbs with their collocations across written academic genres might be followed by advanced foreign and second language users so that their academic writing and publication goals can be accomplished. Since each genre requires certain conventions that each member of this genre is supposed to comply with, learners are also expected to attend this community with full competence. Teachers should help learners gain awareness of the fact that knowledge is socially constructed within particular domains, and thus this line of thinking is reflected to academic writing as well. This basic theoretical background in the minds of teachers can motivate learners to pay attention to certain constructions in a certain genre.

More practically, learners need to be aware of not only common verbs used but more importantly of the collocates each verb attracts because the main competence in writing a professional article in a specific genre requires noticing certain collostructs in this very particular discipline (Hoey, 2001, 2004; Hyland, 2008). Hyland (2008, p.561) suggests that each learner should be trained in a 'genre approach' by teachers who are supposed to regard texts as a dynamic 'social interaction' rather than only a sequence of verbs given in a list. In parallel with this explanation, this study recommends teachers to show the similarities and difference in using collocates. Teachers can direct their attention to specific genres so that they can help learners notice lexico-grammatical patterns in academic writing rather than present a list of verbs or nouns.

This study showed that teaching writing is beyond listing only similar content words because each genre is specially and socially constructed and compromised (Hyland, 2007). It is important for both teachers and learners to discover and develop genre-specific corpora for themselves elaborately, and work on these constructions together. Thus, Hyland (2007, p.251) stresses the fact that 'discursive' similarity as well as variability should be noticed and detected by learners. In this sense, teacher educators should introduce and guide teachers and learners into genre-oriented theory and pedagogy presupposing that learners shall write only in socially constructed domains, and learners should bear in mind that they are liberal and can be creative only within constraints in order to attend the world of socially determined and constructed meanings in academic writing because each genre refers to a particular social world with certain patterns of language (Hoey, 2005; Hyland, 2007). The study has got significant implications for English language learning and teaching, particularly specific to academic writing in that while introducing academic texts to learners, teachers have a reservoir of available data of collocations which they can put into the utilization of language users while producing an academic text. This availability is bound to facilitate the process of writing in general.

In terms of classroom application, teachers and learners have new roles in language teaching and learning because they can constitute their own corpus in the classroom so that they can extract their own collocations and reach reliable generalizations over examples and exemplars. Before learners are asked to write about an academic topic, as a warm-up activity forming a corpus in a two-three week period might prepare learners to use the target language according to the specific topic or genre they are supposed to write. Unless learners are entrenched and enriched by rich data of corpus, deviant forms will be inescapable. Teachers should show learners how to prepare an effective corpus instead of merely giving them hundreds of examples through a concordancer. However, a concordancer can be used to check whether any used collocation in a classroom setting is written or uttered by native speakers. Learners should be able to revisit and recheck the data that they have extracted and studied.

Selective attention of learners may differ from each other in that each learner may attend to different data. Therefore, learners can work together in order to share the data they have chosen during the compilation and selection of lexical collocations (Lewis, 1998). This process will give learners the chance to negotiate the meaning of the data together, which might reinforce learning. By doing so, teachers can give learners the feeling that they are responsible for their own learning, and they learn to be independent while learning a language.

Another implication for ELT is that material writers may have to review their definition of lexical collocations because lexical collocations should also embrace collostructions as well. Material writers should not treat lexis and grammar as separate. Rather, they should show language learners that grammar and lexis can be learnt concurrently (Lewis, 1998; Howarth, 1998). Material writers, in this sense, can help this paradigm change in language learning settings take place. If material preparation contains grammar-lexis activities, then learners will be able to perceive language as holistic and integrative rather than dichotomic.

In terms of testing in ELT, testers should not measure grammar and lexis separately. Rather, they should prepare exams that allow learners to reflect their knowledge of collostructions as well. Since lexical collocations have syntactic functions in language production, it is important to direct learners to focus on these collostructions by developing certain tests containing both collocates and collostructions instead of asking only the meaning of a certain word. It should be borne in mind that each lexeme has its own intrinsic properties that should be perceived by learners. Therefore, testers should gain an awareness of this new paradigm change in language studies.

As a negative implication of this study, it can be said that language is constantly changing, and the data they have collected may change over the years. In addition, being obsessed with fixed expressions may lead learners not to use their creativity in language. Foreign language learners might be able to use their creativity and make contributions to the target language they learn. Therefore, coming up with creative collocations by foreign or second language learners should not be regarded as something negative. Rather, these creative collocations or collostructions should be perceived as a contribution to the field. Instead of labeling these creative collocations as errors, mistakes or deviances, it is better to treat them as possibly acceptable because each new collocation is a candidate to be a part of language. In this sense, language learners should be encouraged to make use of corpus data and to use their own creativity.

## References


Barsalou L.W. (2005). Abstraction as dynamic interpretation in perceptual symbol systems. In In L. Gershkoff-Stowe & D. Rakison (Eds.), *Building object categories* (pp.389-431).Carnegie Symposium Series. Mahwah, NJ: Erlbaum

Bartsch, S. (2004). *Structural and functional properties of collocations in English. A corpus study of lexical and pragmatic constraints on lexical co-occurrence.* Tubingen: Verlag Gunter Narr.

Becher, T. & Trowler, P.R. (2001). *Academic tribes and territories*. Buckingham: SRHE and Open University Press.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243-257.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam : John Benjamins.

Biber, D., Conrad, S. & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. New York: Pearson Education.

Bıckı, A. (2012). *Acquisition of English collocations by adult L2 Turkish learners*. Unpublished doctoral dissertation, Cukurova University, Adana.

Bloom, L. (1973). *One word at a time: The use of single word utterances before syntax*. The Hague,The Netherlands: Mouton

Bloomfield, L. (1933). *Language.* New York: Henry Holt.

Bolinger, D. (1976). Meaning and memory. *Forum Linguisticum I*, 1-14.

Bresnan, J. (1982). *The mental representation of grammatical relations*. Cambridge, Mass: MIT Press.

Carter, R. (1998) *Vocabulary: Applied linguistics perspectives*. London, Routledge.

Charles, M., Pecorari, D., & Hunston, S. (2009). *Academic writing: At the interface of corpus and discourse.* London: Continuum International Publishing Group

Chomsky, N. (1995). *The minimalist program*. Cambridge , Mass : MIT Press

Chomsky, N. (2000). *New horizons in the study of language and mind*. Cambridge : Cambridge University Press