Sudan University of Science and Technology
Deanship of Scientific Research
Journal of Linguistic and Literay Studies

Stylistic Features of Scientific English: A Corpus-based Study

Ahmad Muhammad Atiya
Karary University - College Of Languages

Abstract:
The current study aims at investigating the stylistic features of scientific English. An analytic, descriptive research design was adopted. The qualitative data were collected from the British National Corpus where 11 files associated with scientific English were generated. The findings have indicated that the scientific English is associated with some stylistic features such as the use of the present simple, the passive, the prefixes and suffixes. The findings have also revealed the frequent use of abstract nouns and numerical expressions in scientific English which seem to be free from personalized expressions.
Keywords: *stylistic, corpus, files, qualitative, -feature*

المستخلص :
تهدف هذه الدراسة إلى استقصاء السمات الأسلوبية المتعلقة باللغة الإنجليزية العلمية. فقد تم استخدام منهج البحثي التحليلي-الوصفي. حيث تم الحصول على المعلومات النوعية من المخزونة البريطانية. فقد تم جلب أحد عشر ملفا متعلقا باللغة الإنجليزية العلمية. أشارت نتائج الدراسة إلى أن الميزات الأسلوبية المتعلقة باللغة الإنجليزية العلمية تتمثل في استحدام المضارع البسيط ، و المبني للمجهول ، و السوابق و اللواحق . و قد أشارت النتائج أيضا إلى الاستخدام المتكرر للأسماء و التعابير الرقمية في اللغة الإنجليزية العلمية و التي تبدو خالية من التعابير ذات الطابع الشخصي.
كلمات مفتاحية : *أسلوبي ، مخزونة لغوية ، ملفات ، نوعي ، ميزة.*

Introduction:

Scientific English has proven to be different from other language genres in terms of nature and function. It is precise, accurate and impersonal. It is an objective interpretation of facts and findings. It is based on scientific experiments rather than on human assumptions. Additionally, scientific English tends to be free from flowery or aesthetic language (Close, R. 1965). It is reported that scientists focus sharply on the authenticity of the theme rather than on the fashion of demonstration. Thus, there is no room for human impulse and human pleasure in scientific articles (Ding, D.2002). Lexically speaking, the scientific English vocabulary includes a wide range of content words and functional words. The high frequency of content words is due to the fact that scientists usually deal with concrete objects and substances. Functional words such as prepositions and articles are used more frequently in the scientific genre to identify objects or people. In the same context, the vocabulary of scientific English includes (technical terms) which are invented to define new phenomena and to explain new things and processes. Each scientific subject has its store of terms with precise, narrow meanings. Examples of terms related to the field of biology are, photosynthesis, phylum, chlorophyta, agents, species, fertility, algae, vegetation, gametes, zygote, vesicle, host, cyanobacteria. chlamedomonas. In addition to

technical terms, sub-technical terms are common in scientific English. Sub-technical terms are composed of the words which are not specific to a subject speciality but which occur regularly in scientific and technical texts, e.g., reflection, tendency, isolation, and density. They form about 457 (76%) of the total number of nouns in the British National Corpus . These will have priority in the language programme. They are commonly met in general English but they take a specialized meaning within a scientific and technical context, e.g., cycle (its use in blood cycle). Additionally, compound words are commonly used in scientific English. This is related to the way of scientific thinking because a scientist usually tends to express his/her ideas accurately and in a brief condensed way. So, instead of saying: transmission of virus by seed; he/she will say virus seed transmission. Similarly, a disease which is caused by a fungus is a fungus disease and a tube used for performing tests is a test-tube.

Thus, the current study aims at examining the stylistic features of scientific English. The primary objective of this examination is to help lexicographers, teachers, students and researchers have insight into the scientific genre. In this way they are able to analyse and interpret the scientific genre. In this study, the researcher hypothesizes that scientific English is oriented to the inclusion of affixation, present simple, passive, abstract nouns and tends to be detached from impersonalized language. To confirm these hypotheses, several files of scientific language will be generated from the British National Corpus to investigate the range of frequency of use concerning the features assumed-above.

Literature Review

The literature concerning this study is divided into two parts: (1) conceptual framework and (2) review of previous studies.

Part One: Conceptual Framework

In surveying the literature related to the field of scientific English, we can find a body of research has addressed this issue. For instance, Stevens ((1977) pointed out that the scientific discourse uses several suffixes and prefixes oriented to Latin and Greek origin. In addition, Ewer (1971) found out that the scientific language is concerned with technical meaning, suffixes and prefixes and qualifying and quantifying words and phrases. Widdowson (1974) pointed out that scientific language avoids the first and the second person.According to Robert, A, Day and Barbara Gastel (2011) 'Scientific writing is the transmission of clear signal to a recipient. Scientific writing needs no ornamentation. Flowery literary embellishment –metaphor, similes, and idiomatic expression are very likely to cause confusion and should seldom be used in research paper". Thus, all scientists should use English language with precision which is extremely instrumental in communicating the scientific findings (Day, R.A. & Sakaduski, N. D, 2011). Concerning the type of grammar used in scientific language, Hilary Glasman – Deal (2009) found out that the present simple tense is commonly utilized to state accepted facts and truth. In addition, the passive is similarly used to highlight the most important participation or events within the context. It is also argued that many scientists and engineers believe that everything should be written in passive voice (Gilbert 1976; Vande Kopple 1994; Hyland 1996; Ding 2002; Dorgeloh. H. 2004).

Part Two: Review of Previous Studies

- ***Corpora and Corpus Linguistics***

This section is devoted to corpora plus corpus linguistics. These two elements will be discussed in terms of their ontology, concerns, techniques, types and effectiveness.

• ***. Definition of Corpus Linguistics***

According to McEnery and Wilson (2001, p.1), corpus linguistics is "*the study of language based on examples of "real life" language use*". Similarly, Aijmer and Altenberg (1991) explained that corpus linguistics is the study of language on the basis of text corpora.

Based on the previous definitions, it is clear that corpora have widely been used in analysis and study of language because they have a broader spectrum of representativeness. This merit has qualified corpora to widely be used in several subfields of Applied Linguistics: lexicography, grammar, socio-linguistics, translation, language learning and teaching, stylistic analysis, dialectology and historical linguistics.

• ***Concerns of Corpus Linguistics***

Leech (1992) stated that key concerns of corpus linguistics should focus on:

I. linguistic performance, rather than competence;
II. linguistic description; rather than linguistics universals;
III. quantitative; as well as qualitative analyses;
IV. a more empiricist; rather rationalist view of scientific inquiry.

In the same way, Biber (1998) identified four features for the utilization of corpus linguistics in the analysis of a language in terms of being:

V. empirical, analyzing the actual patterns of use in natural texts;
VI. utilizing a large and principled collection of natural texts, known as a "corpus", as the basis for the analysis;
VII. making extensive use of computers for the analysis, using both automatic and interactive techniques.

Therefore, in the present study both quantitative and qualitative data are utilized.

Corpus Linguistics Techniques

O'Keeffe, Carter and McCarthy (2007) stated that researchers can make use of certain techniques on the corpus, utilizing standard software such as *WorldSmith Tools* and *Monoconc Pro*. Such techniques are projected in: concordancing, word count (word frequency), key- word analysis, cluster analysis and lexico-grammatical profiles. With regard to Concordancing, it is considered as a key tool in the corpus search and it means using corpus software to find every instance of a particular word or a phrase. The search word or phrase is often referred to as the "node" and the concordance lines are usually presented with node word/phrase in the centre of the line with seven or eight words presented on either side. These are known as Key- word -in Context displays (or KWIC concordances). In the current study, concordance lines will be used to show words instances under investigation.

Another common corpus tool which software can generate is the rapid calculation of word frequency for any batch of a text. The value of this type of search is to facilitate enquiry across various corpora, different language varieties and various contexts of use. Therefore, such a technique is of great importance to our present study because we rely on the interpretation of frequency and data distribution to determine lexical items that are common and those which are uncommon, and hence helps infer the meaning of some vocabulary. Additionally, key- Word analysis is one of the most striking corpus techniques because it allows the identification of the key- words in one or more

texts. According to Scott (1991), key- words are those whose frequency is remarkably high compared to some norms. It is noted that the key- word provides a useful way of describing a text or a genre and has potential application in the area of forensic linguistics, stylistics, content analysis and text retrieval.

Cluster analysis, on the other hand, is useful in the analysis of the systematic combination of words or "chunks" (e.g. *I mean*). This technique can provide insights into vocabulary description, teaching and acquisition. So, the researcher will make use of such a technique when applying vocabulary learning strategies.

Finally, the technique of Lexico-grammatical profiles has remarkably distinguished itself in the analyses of corpora. This is due to the fact that concordance lines can provide the researchers with the lexico-grammatical profiles of a word and its context accompanied by its collocations, chunks/idioms, syntactic restrictions (e.g., prepositions use, typical clause-positions and tense-aspect) and semantic restrictions (e.g., words or phrases that are applied to human only).

Concept of Corpora

According to Oxford Advanced Learner's Dictionary (2013), a corpus is a collection of written or spoken texts. From the linguistic perspective, a corpus is a large amount of language data stored on a computer for the purpose of linguistics analysis. McEnery and Wilson (1961, p.24) defined a corpus as "a *finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration*". Similarly, Gries, S. (2004,p.7) defined corpora as '*a machine-readable collection of (spoken or written)texts that were produced in a natural communicative setting, and the collection of texts is compiled with the intention (1) to be representative and balanced with respect to a particular linguistic variety or register or genre and (2) to be analyzed linguistically*'.

Types of Corpora

Bibber et al., (2004) argued that a corpus is designed according to the search type that is going to be addressed. Therefore, corpora are classified into eight types. One of the broadest types of corpora is a generalized corpus. The generalized corpus is often very large, more than 10 million words, and contains a variety of language so that findings from it may be somewhat generalized. Although no corpus will ever represent all possible language, the generalized corpus seeks to give users as much of a whole picture of a language as possible. The British National Corpus (BNC), the American National Corpus (ANC) and Corpus of Contemporary American English (COCA) are examples of generalized corpora. The second type of corpora includes the Specialized Corpora. It is a specialized corpus contains texts of a certain type and aims to be representative of the language of this type. Specialized corpora can be large or small and are often created to answer very specific questions. Examples of specialized corpora include the Michigan Corpus of Academic Spoken English (MICASE) and the CHILDES Corpus which contains language used by children. Paradoxically, a learner corpus, on the other hand, is a kind of specialized corpus that contains written texts and / or spoken transcripts of a language used by students who are currently acquiring the language. A well-known learner corpus is the International Corpus of Learner English (ICLE) which is often tagged. Besides the aforementioned corpuses, a pedagogic corpus has made a difference mainly in the field of teaching and

learning. It is a corpus that contains language used in classroom settings. The Pedagogic Corpus can include academic textbooks, transcripts of classroom interactions, or any other written texts or spoken transcripts that learners encounter in an educational setting.Again, one of the techniques used to describe languages is comparing them. In this spirit, comparable corpuses are created to obtain such a goal. They are used to compare corpora from various languages such as English and Spanish or various varieties of a language. The sixth type of corpora is the parallel corpora. They comprise two or more corpora in different languages, each including texts that have been translated from one language into another. They can be used by translators and by language learners to discover the potential equivalent expressions in each language and to investigate differences between languages. The seventh type of corpora is the diachronic corpora. They include texts from various periods of time. They are used to trace the development of aspects of a language over time. Good examples of such corpora include the ARCHER **(**A Representative Corpus of Historical English Registers) and Helsinki Corpus. Finally, the monitor corpora are utilized to trace the current changes in a language.

The Use of corpora in Pedagogy

Leech (1997) stated that the interplay between corpora and pedagogy focuses on three areas: indirect use of corpora, direct use of corpora and teaching-oriented corpus development. With regard to the indirect use of corpora in pedagogy, it is true to say that a corpus is playing a major role in reference publishing, syllabus and materials development, language testing and teacher development. With regard to reference publishing, publishers can make use of taggers and frequency information in reference publishing.

Nowadays, many scholars have entirely made the maximum use of corpora to critically look at some learning material such as TOEFL (Teaching of English as a Foreign Language), syllabuses and teaching materials. They actually depend on the huge data which a corpus provides. Teacher development, on the other hand, has become the key focus of a corpus. Benefitting from the huge packages of information that a corpus provides, a corpus is used to raise the language awareness of English teachers. Moreover, corpora are used in language testing. Kaszubski and Wojnowska (2003) revealed that some annotated (coded) corpora have recently been used: as an archive of examination scripts; to develop test materials; to optimize test procedures; to improve the quality of test marking; to validate tests; and to standardize tests. With regard to the direct use of corpora, they have become rather a distinct source of information in teaching compared to the traditional teaching methodologies. McEnery and Wilson (2001) indicated that the main scope of corpora in learning is interacting, inducting and illustrating. Illustration' means looking at real data, 'interaction' means discussing and sharing opinions and observations, and 'induction' means making one's own rule for a particular feature, which will be refined and developed as more and more data is encountered. In contrast, the traditional teaching methods focus on practicing, producing and presenting of information. The third type of the uses of corpora in the area of pedagogy is teaching-oriented corpus development. This technique is particularly useful in teaching languages for specific purposes (LSP corpora) and in research on L1

(developmental corpora) and L2 (learner corpora) acquisition Leech (1997).

In the above section, a detailed account is about corpus linguistics and corpora. This description has dealt with their concept, types, and techniques besides showing the efficacy of corpora to the present study.

Corpus-based Approach vs. Corpus-driven Approach

According to Biber (1998), corpus studies have used two major research approaches: corpus-based and corpus-driven. Corpus-based research assumes the validity of linguistic forms and structures derived from linguistic theory. The primary goal of research is to analyze the systematic patterns of variation and use for those pre-defined linguistic features. Corpus-driven research is more inductive, so that the linguistic constructs themselves emerge from analysis of a corpus.

Materials and Methods

In the current study, a qualitative design will be adopted. That is, 100 files addressing scientific genres will be generated from the British National Corpus. Then the data will be described linguistically in light of Swale's (1990) model. Swale's model is utilized here to explore how scientific facts and findings are intrinsically formulated and what kinds of linguistic features govern varied segments of scientific research articles.

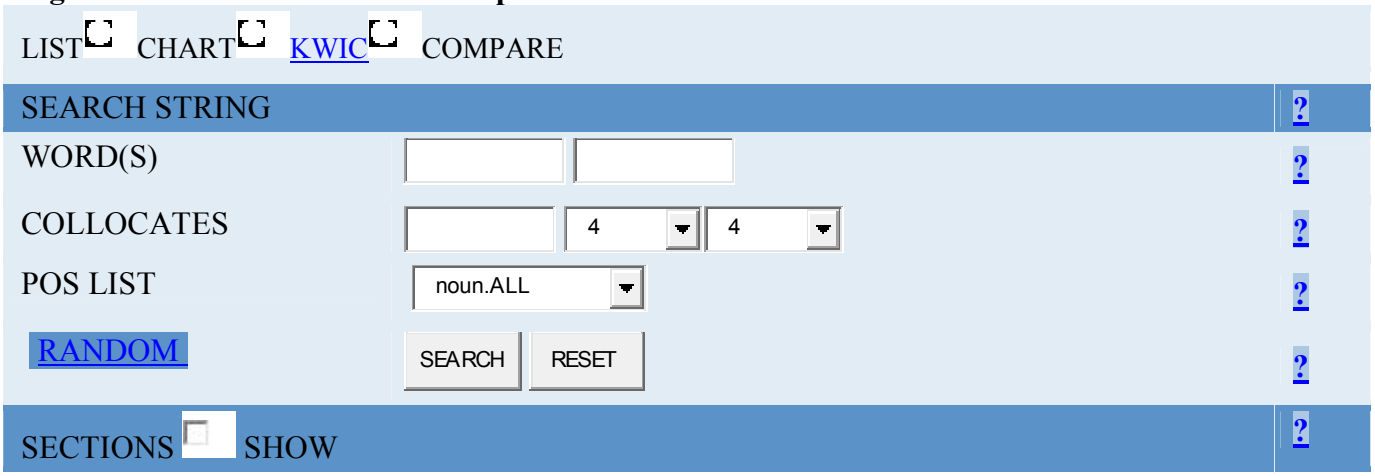
Sampling

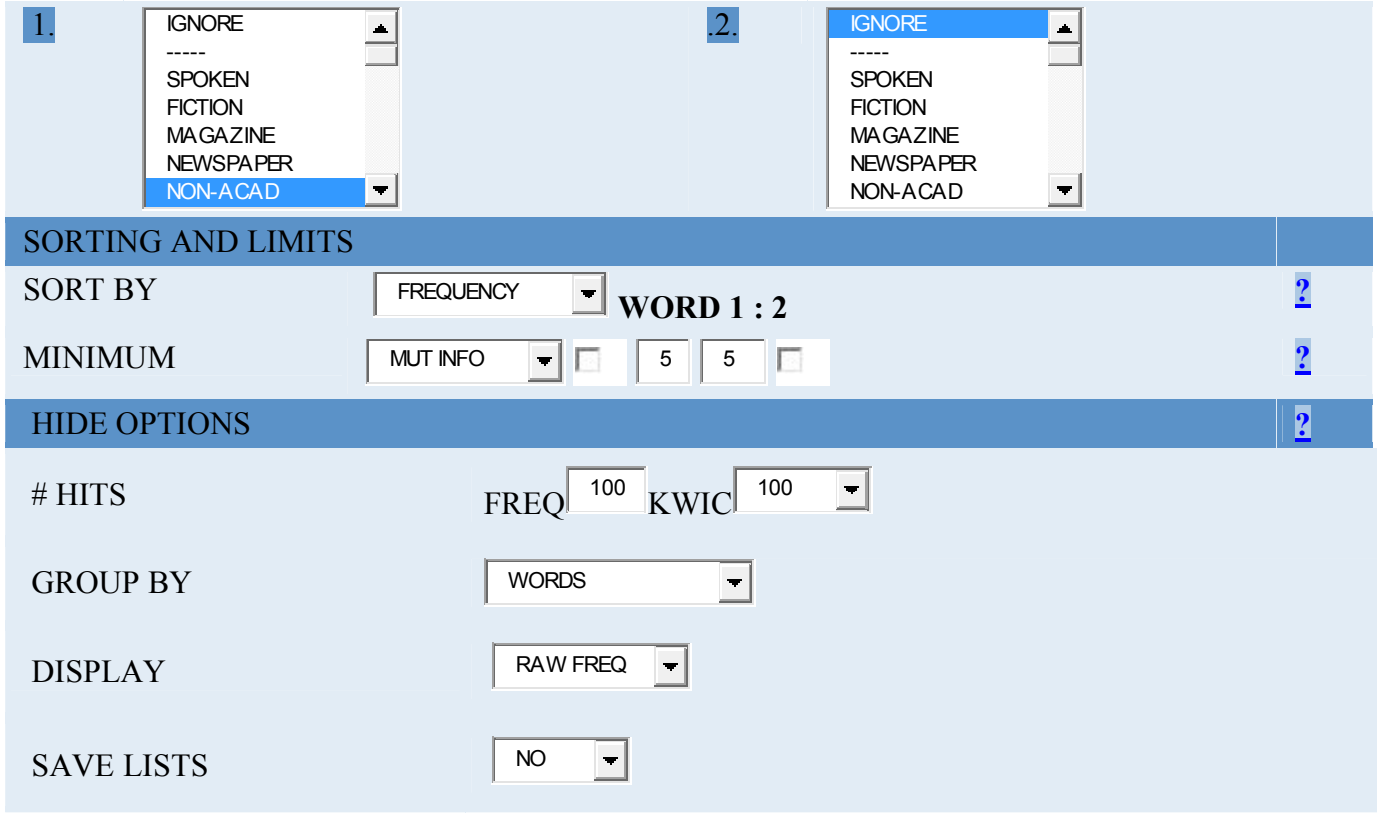
. The ample size was 11 files associated with scientific English will be generated. The sample was drawn from the British National Corpus which was created by Oxford University, Lancaster University and the British Library between 1991 and 1994. It incorporates a wide spectrum of written and spoken texts in English. It is a hundred-million-word corpus

Procedures of Data Collection

To collect the desired data, the researcher made use of the corpus interface (see figure 1). First, the type of genre (scientific genre) was selected. Second, KWIC (key words in context) was clicked. Then packages of information in forms of files were displayed. Finally, the number of features in each category was counted and analyzed qualitatively.

Figure 1: The British National Corpus user's interface





Results of Data Analysis

File 1: Every substance melts and boils……

File 1 shows the use of the present simple tense. (melts)

File 2: Most solids can't be squeezed into ……

File 3: A 1.5 cell is used with a 3 V buzzer …

File 4: Some materials can be recycled…..

In file 2,3 and 4, the use of the passive is projected.(can be recycled, be squeezed).

File 5: The earth is home to more than seven billion people. It is also home to billions of animals and plants.

File 5 addresses the inclusion of numerical expression in scientific genre (billions of animals and plants , seven billion people).

File 6: Before scientists can make a key, they have to observe the animals very carefully. One key is the environment ….habitat…..

File 7: Snails. They have good eyesight. They can move quickly and have strong beaks. The movement

File 6 and 7 shows that adjective and adverbs are commonly used.

File 8: Here is an identification key…..

File 9: Symptoms are the signs of illness

File 10: Garbage and litter should be recycled.

File 8, 9 and 10 display the use of affixation (-ness. –tion- ,re-, -ment)

File 10: ….the pull of earth's gravity…..

Files 10 shows the use of abstract nouns (gravity).

File 11:It is reported that amphibians live in …….

File 11 is detached from the use of the first and second persons.

Discussion

The current study aims at analyzing the stylistic features regarding scientific English. Two hypotheses are addressed: (1) scientific English is associated with some linguistic features such as the inclusion of affixation, present simple, abstract nouns, passive and numerical expressions. It is hypothesized also that scientific English is free from personalized expressions. The results of the analysis of the files generated from the corpus have proven that scientific English utilizes the present simple tense and passive. This fact is congruent with Hilary Glasman – Deal (2009). I think the inclusion of the present simple tense is due to the fact that science addresses facts and truth and it does not rely on hypotheses or on assumptions. In addition, the results have also shown that scientific English is oriented to the use of affixation. This fact is also reached by Ewer (1971) and Stevens ((1977). I think the inclusion of affixation in scientific genre is essential because scientific English is rich in abstract and concrete nouns.

Furthermore, the findings have revealed that scientific English is rich in numerical expressions. This result is similarly reached by Ewer (1971). The researcher views that the inclusion of numerical expression might be attributed to the fact that scientific English tends to provide calculations and statistical information so as to provide precise facts. The findings have also revealed that the scientific writing incorporate the use of qualifying phrases and use of passive. This finding is congruent with Hilary Glasman – Deal (2009). An interesting interpretation of this feature is that science always tries to transmit a living image to the audience. It also tries to make a sharp focus on the events and actions being described. Moreover, the findings have also indicated that the scientific English is free from the use of first and second persons. The same result is reached by Widdowson (1974). This could be interpreted by that scientific English is always free from personal impulse and depends heavily on the description of the universe.

Findings

In light of the analyses of the data, the following findings have been reached:

1- Scientific English always includes the use of present simple.

2- Scientific English involves the use of passive voice.

3- Scientific English uses affixation.

4- Scientific English uses qualifying phrases and numerical expressions.

5- Scientific English is detached from personalized language.

6- Scientific English is rich in abstract nouns.

Conclusion

The main objective of the current study is to highlight the key stylistic features concerning scientific English. The findings that have been reached are expected to distinguish scientific English from other types of writing. The findings also indicate that scientific English is largely based on precision, objectivity and authenticity. The findings have also suggested that further studies on other genres of scientific English should be conducted so as to identify the most frequent and the least frequent stylistic features among them.

References

1. Biber, D., Conrad, S. And Reppen , A. (1998) *Corpus Linguistics: Investigating Language*

2. *Structure and Use*. Cambridge: Cambridge University Press.

3. British National Corpus (1991). [Online]. Available from: http: //www. nat

4. corp.ox.ac.uk/ [Accessed: 11/4 /2016]

Close, R. (1965). *English We Use for Science*.London, Longman.
5. Day, R. A., Sakaduski. N. D. (2011). *Scientific English: A Guide for Scientists and Other*
6. *Professionals.* (3rd ed) Greenwood. p. 4. [Online] Available: http://www.abc-clio.com/product.aspx?id=2147491826
7. Ding, D. (2002). *The passive voice and the social values in science. Journal of Technical Writing*
a. *and Communication*
8. Dorgeloh. H. (2004). *The limits of variation in scientific writing: Syntactic and functional*
9. *constrains. I T L. Review of Applied Linguistics,* 143/144, 199-222. http://dx.doi.org/10.1177/0741088307302946

16. Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge:
17. Cambridge University Press.
18. Kennedy, G. (1998) *An Introduction to Corpus Linguistics*. Harlow: Longman.
19. Leech, G. (1997) '*Teaching and language corpora: a convergence' in A*. Wichmann, S.
20. Fligelstone, T. McEnery and G. Knowles (eds.)
25. Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford
a. University Press.
26. Stevens. P. (1977). *Special Purpose Language learning: a perspective*. *Language Teaching*
27. *&Linguistic Abstracts*, 10(3), 145-163. http://dx.doi.org/10.1017/S0261444800003402
28. Swales. (1990). Genre analysis: English in academic and research settings. *English for Specific*
29. *Purposes*, 20,439-458

10. Ewer J. R. (1971*). Further notes on developing an English programme for students of science*
and technology. English. *Language Teaching Journal*, 26(1), pp. 65-70.
11. http://dx.doi.org/10.1093/elt/XXVI.3.269
12. Gilbert, G. N. (1976). *The transformation of research findings into scientific knowledge*. *Social*
13. *Studies of Science,* 6,281-306. http://dx.doi.org/10.1177/030631277600600302
14. Gries, S. (2009). *Quantitative Corpus Linguistics with R: A practical Introduction.* Routledge: New York.
15. Hyland, K. (1994). *Hedging in academic writing and EAP textbooks*. *English for Specific Purposes*. http://dx.doi.org/10.1016/0889-4906(94)90004-3
21. McEnery, T. and Wilson, A. (2001) *Corpus Linguistics* (2nd edition).[e-book] Edinburgh: Edinburgh University Press.
22. O'Keeffe, A., McCarthy, M., Carter, R., (2007). *From corpus to*
23. *classroom: Language use and language teaching.* Cambridge
24. University Pres.

30. Vande Kopple, W. J. (1994). *Some characteristics and functions of grammatical subjects in*
31. *scientific discourse.Written Communication*, 11, 534-564. http://dx.doi.org/10.1177/0741088394011004004
32. Widdowson H. G. (1974). *Literary and scientific uses of English*. *English Language Teaching*
33. *Journal*, 28(3), 282-292. http://dx.doi.org/10.1016/0889-4906(91)90015-O