A Comparative Analysis of Lexical Bundles in Journalistic Writing in English and Persian: A Contrastive Linguistic Perspective

Marzieh Rafiee, Ph.D Student, Department of English, Isfahan University, Isfahan
rafieemarzieh@gmail.com

Mahbube Keihaniyan, MA Student, Young Researchers and Elite Club, Najafabad Branch, Islamic Azad University, Najafabad, Isfahan
m_keihaniyan@yahoo.com

Abstract
This paper investigates the use of 'lexical bundles' in two broad corpora of journalistic writing. The aim of this study is to compare the use of lexical bundles in the two domains, one consisted of newspaper articles written in English and published in England and the other one comprised of newspaper articles written in Persian from Iranian publications. For this purpose, the frequency of occurrence and distribution of different functional taxonomies of lexical bundles across the subject matter were investigated. More than 2.5 million words of different English and Persian-produced online newspapers were collected and they were identified by the help of two computer programs, then their functionswere analyzed. Consistent with similar research on lexical bundles, the analysis indicates that most bundles perform a referential function in journalistic register. These findings may be particularly useful to translators and also EFL practitioners, as they seem to give new insights into the development of learner language.

Keywords: ESP; lexical bundles; journalistic writing; newspaper register

During the last century, the study of word combinations has attracted many linguists and researchers. What made researchers more interested, was the use ofthese building blocks by EFL learners of English.In a study done by Milton (1998)the essays written by Hong Kong students and native English speakers werecompared and it was concluded that Hong Kongstudents used more recurrent wordcombinations, compared to their counterpart, native speakers.

Lexical bundles, a particular and relatively newborn category of word combinations, are words which follow each other more frequently than expected by chance, helping to shape text meanings and contributing to our sense of distinctiveness in a register. These units of language have shown that language is "register specific and perform a variety of discourse functions" (Allen, 2009, p.367). Therefore, as Haswel (1991) states, the application of these fixed expressions indicates the proficiency level and the success of language learners in that specific register. Furthermore, learning to use the more frequent fixed phrases of a discipline can contribute to gaining communicative competence in a field of study, there are advantages in identifying these clusters to better help learners acquire the specific rhetorical practices of their communities (Hyland, 2008).

Research on lexical bundles has encompassed both on spoken and written prose (Biber& Barieri2007;Hyland, 2008). The results show that lexical bundlesdiffer across these two registers in terms of structure and function. Structurally most bundles in speaking are classified as clausal which consist of a verb phrase fragments while bundles in written prose are mostly phrasal which contain noun phrase or prepositional phrase fragments. On the other hand, bundles play stance roles (expressions like *I know that, it is important that, let's have a look*) in spoken register but

language learners apply more referential bundles (expressions like *in the United States, at the same time*) in their writing (Cortes, 2008). In EFL contexts, such as Iranin which the development of written skills is more dominant than spoken ones, the language learners need to know the functions and structures of these fixed expressions to be able to improve their writing skill.

Regardless of those studies done in French (Salem, 1987) and Spanish (Cortes, 2008) the focus of research on the recurrent lexical sequences was mostly on English registers (Cortes, 2008). Johansson (2007) discussed the possibilities as well as limitations of multilingual corpora in linguistic research. Stating that the prediction of potential problems for language learners on the basis of contrastive analysis is only one step, Johansson (2007) believes that one of the merits of multilingual corpora is to provide insightful details of thedata of learner language use and the languages involved in the comparison. Johansson (2007) also emphasizes on the relationship between corpora and translation training and language teaching, arguing that parallel corpora provides fertile grounds for both fields. Also De Cock (2000), in her study on essays produced by English and French EFL learners, states French EFL learners used more word combinations than native speakers of English. The studies done on lexical bundles, as a new category of word combinations, more focused on L1 speakers' production of lexical bundles in both conversational and academic registers. The example studies are the ones conducted by Biber and Conrad (1999) who analyzed the use of lexical bundles in academic writing and conversation, Hewings&Hewings (2002), who compared the use of lexical bundles in the written production of published authors and university students, Cortes (2002a, 2004), who identified four-word lexical bundles (called target bundles) used by published authors in history and biology and by students at three different levels in those disciplines, Biber, Conrad & Cortes (2004), who described the use of lexical bundles in two university instructional registers: classroom teaching and textbooks, Biber&Barbieri (2007), who investigated the use of lexical bundles in a wide range of spoken and written university registers, including both instructional registers and students advising/management registers (e.g., office hours, class management talk, written syllabi, etc.), and finally Hyland (2008), who explored forms, functions and structures of lexical bundles in three disciplinary variation; research articles, doctoral dissertations and Master's theses. Although corpus-based investigations of natural language data have established the existence of frequent recurring multi-word lexical chunks in texts (Biber et al. 1999; Cortes2004; Sinclair 1991; Stubbs & Barth 2003), there is still disagreement on their definition and classification. The major controversy has been that it is hard to define them in an unambiguous manner. Moreover, researchers do not often agree on the way they should be classified, which has resulted in a multitude of taxonomies: "lexical bundles" (Biber et al. 1999), "prefabs" or "lexical phrases" (Nattinger & DeCarrico 1992), "formulaic sequences" (Schmitt & Carter 2004), "sentence stems" and"clusters" (Scott1996). By whatever name they may be called , the common denominator seems to be that lexical bundles are text – generated ; and may be any combination of word (s) occurring together most frequently in a given register (Biber 2006) . In this present study the researchers prefer to use the term" lexical bundle", since it is the most widely used in the literature.

The Present Study

It is possible to hypothesize that investigating lexical bundles in written language produced by native English speakers and native Persian speakers throw more light on the different quality of their written language. Therefore, the purpose of this study is to investigate and analyze the frequency and distribution of different functional taxonomies of lexical bundles

across English and Persian-produced online newspapers to find some similarities between these two corpora.

Materials

Corpus used for the study. The present study is based on an analysis of different parts of newspapers (e.g., Domestic Economy, World, Art & Culture, Middles East, Politics and Science, etc.). The texts used in this corpus belong to four newspapers, two of them published in Iran (Iran and Etela't) and the other ones published in England (Times and Independent) in 2010 and 2011 issues of the newspapers. The newspapers were chosen as the source of corpus collection because they were online and accessible to download the necessary texts. Besides, they were more popular than other English newspapers in Iran and England in terms of readership. At least six parts in each newspaper were selected because they contained more words than the other parts of the newspapers in each number. More than 2.5 million words of these parts were collected and the lexical bundles were identified by the help of two computer programs: Antconc3.2.1w (Anthony, 2007), and Wordsmith tools5 (Scott, 2008). The former was used for identification of lexical bundles and concordancing while the latter was only used to find the number of texts within which each bundle had been used.

Bundles identification.As Biber, D., Conrad, S., & Cortes, V. (2004) state in their study on the bundles, frequency has the key role in identification of bundles. "… Frequency data identifies patterns that must be explained." (p. 376). Although the actual frequency cut-off point used by different researchers is arbitrary, in the present study, the cut-off point 10 times in a million words for each of the two corpora was selected. For English texts, Antconc3.2.1w (Anthony, 2007) computer program was used to explore lexical bundles, their frequencies, the number of texts in which they had been used, and their actual contexts of use. In the case of Persian corpus, because there was no computer program to find the frequency of bundles, this process was doneby the help of Microsoft Word. For this purpose, all the non-textual production of the Persian texts (page numbers, references, etc.) were removed and the Microsoft Word browsed through the texts to find the words that were supposed to be bundles, then the frequency of the bundles found were calculated manually by the researchers. In this study like some other previous studies of lexical bundles (e.g. Cortes, 2002), only four-word combinations of bundles were investigated in English texts. For Persian texts, because the occurrence of three-word combinations of bundles was prevailing, they were identified as lexical bundles. When all the texts had been processed, all the bundles which were repeated at least 10 times in more than 2.5 million words and in more than ten texts in each of these two corpora were treated as lexical bundles.

Results and Discussion

Using the revised version of the functional taxonomy, lexical bundles were classified according to the functions they performed in newspaper register. Because some bundles were not present in the corpuses used in previous studies, some new subcategories of bundles need to be created in order to classify the rest of bundles found in newspaper corpus. The findings here, once again, confirm that the *referential* category of lexical bundles prevailsthe parallel corpora. Table 1 below reveals the functional classification and identification of bundles in the analyzed texts. Biber, Conrad, and Cortes (2004) fully describe the main discourse functions and of the sub-categories of discourse functions identified within them.

Table 1. *English lexical bundles classified functionally*

Categories	**Subcategories**	**Lexical Bundles**
Referential bundles	Identification/focus	Is one of the, one of the most, as one of the, one of the best, is the first time, to be one of, for the first time, it was the first,
	Specifying attributes *Quantity specification*(1) *Tangible framing attributes*(2) *Intangible framing attributes(3)*	As a result of(3), hundreds of thousands of(1), in the case of(3), in the form of(2), the size of the(2), in a series of(3), a great deal of(1), a lot of people(1), the state of the(3)
	Reference bundles *Time(1)* *Place(2)* *Multi-functional(3)*	The end of the(3), the beginning of the(3), in the middle of(3(, the time of the(1), for a long time(1), the top of the(3), on the verge of(1), in the aftermath of(1), at the beginning of(3), for the first time(1), in the United States(2), in the Middle East(2), at the same time(1), all over the world(2), over the course of(1), the rest of the(3),
Discourse organizers	Topic introduction/focus	in the aftermath of, said in a statement, the first time that, the first time since,
	Topic elaboration/clarification	for the sake of, in terms of the, in addition to the, on the other hand, as well as the, he added that the, in a way that, with the help of, as part of the, in the face of, that it would be
Stance expressions	Attitudinal *Desire(1)/Obligation(2)/Intention/predictions(3)* *Ability(4)Instrumentality(5)*	In a bid to(1&2), is going to be (3), to set up a(3), in an attempt to(4), in charge of the(2), to be able to(4), is expected to be(3&1), is likely to be(3), not be able to(4), to deal with the(5), There will be no (3)
	Epistemic stance *Personal(1)* *Impersonal(2)*	I think it is(1), the fact that the(2), of the fact that(2), is believed to be(1),

Table 2. *Persian lexical bundles classified functionally*

Categories	Subcategories	Lexical Bundles
Referential bundles	Identification/focus	yikÓāzbihtārin, bi unvāniyikÓ, yikÓāzānhā, ,bāyikÓāz یکی از بهترین،به عنوان یکی، یکی از آنها، با یکی از،
	Specifying attributes (۱)*Quantity specification* (۲)*Tangible framing attributes* *(۳)Intangible framing attributes*	Barkhidigārāz(1), t'dādiziydiāz (1), bi Ónşūrāt (2), dārhālipishrāft (3), dārÓnşūrāt (1), dārdāstūrikār (3), bāhuzūrdār (2), dārmubārizibā (۲) (3), dārmuqābilibā (3), dārhāliānjām برخی دیگر از(۱)،تعداد زیادی از(۱) ، به این صورت(۲)، در حال پیشرفت(۳)، در این صورت(۱)، در دستور کار(۳)، با حضور در(۲)، در مبارزه با(۳)، در مقابله با(۳)، در حال انجام(۲)
	Reference bundles *Time(1)Place(2)Multi-* *(۳)functional*	Dārhālihāzir (1), dārsāliguzāshti (1), pāsāzān (3), bā'dāzān (3), dārsārāsārikishvār (2), dārjāmikhābārnigārān (2), dārchāndsāl (1), dārānzāmān (1), dārbārkhikishvārhā (2), hār chi zūdtār (1), tā bi hāl (1), āmmāpāsāz (3), ghāblāzān (3), kipāsāzān (3), ghārārgiriftiāst (3), (۱) tāpāyānisāl, bārāyinukhustinbār در حال حاضر(۱)،در سال گذشته(۱)،پس از آن(۳) ،بعد از آن(۳)، ، در سراسر کشور(۲)، در جمع خبرنگاران(۲)، در چند سال(۱)، در آن زمان(۱)، در برخی کشورها(۲)، هر چه زودتر(۱)، تا به حال(۱)، اما پس از(۳)، قبل از آن(۳) که پس از(۳)، قرار گرفته است(۳)، تا پایان سال، برای نخستین بار(۱)
Discourse organizers	Topic introduction/focus	y, bātāvājoh bi, bābāyāniÓnki, Ćbi guftiyi v ,bāi'lāmiÓnki به گفته ی وی، با توجه به، با بیان اینکه، با اعلام اینکه،
	Topic elaboration/clarification	Dar ḥālika, daradāmehbā, baashārehba, darpāsukhba, darrābatahbā, kadarān, azānba, ānrāba,baharḥāl,kabaunvāna, ānastka,bāanteqadaz, alāmkardka, azṭarafadigar,darhaminkhuṣuṣ, darṣuratika در حالی که، در گفتگو با، از آنجا که، در ادامه با، با اشاره به، در پاسخ به، در رابطه با، که در آن،از آن به ، آن را به، به هر حال، که به عنوان، آن است که، با انتقاد از، اعلام کرد که، از طرف دیگر، در همین خصوص، در صورتی که
Stance expressions	Attitudinal *(۲)Desire(1)Intention/predictions* *(۳)Instrumentality*	Dar naẓardārad(1), anjāmkhāhadshud(2), qarārastbā, kamumkanast (2),agar qarārbāshad(2), darnaẓargarafta(1), muvājahshudaast (3), kaqarārast (2), vaqarārast (3), qarārastdar (3), ṣuratgaraftahast (3),

		,(۳)ṣuratgaraftahdar, bāastafādaaz در نظر دارد(۱) ، انجام خواهد شد(۲)، قرار است با(۲)، که ممکن است(۲)، اگر قرار باشد(۲)، در نظر گرفته(۱)، مواجه شده است(۳)، که قرار است(۲)،و قرار است(۳)، قرار است در(۳)،صورت گرفته است(۳)، صورت گرفته در(۳)، با استفاده از(۳)،
	Epistemic stance *(۱)Personal* *(۲)Impersonal*	Ba naẓar man (1), baataqād man (1), muham in (۲) ast (2), mutaqadastka (2),zamntakid bar به نظر من(۱)، به اعتقاد من(۱)، مهم این است(۲)، معتقد است که(۲)، ضمن تأکید بر(۲)

As it is shown in the tables above, the two corpora have certain similarities. The most obvious of them is that proportionately newspaper writing in English and Persian is dominated by referential bundles compared to the other two categories of bundles.Also, Kjellmer (1994) found that referential bundles in English newspapers are more than other bundles. As regards the other subcategories of bundles, the results indicate that there is equivalent proportionate between the numbers of bundles below these subcategories. In line withfindingsof previous studieslike Yorio (1980) English newspaper register applies topic elaboration/clarification more than the other subcategory, i.e. topic introduction/focus, this is the same in Persian journalistic writing. Regarding the "equivalent and quasi-equivalent" bundles "Cortes (2008), it is concluded that only 12 percent of bundles in English have an equivalent form in Persian corpus. Cortes (2008) defines equivalent bundles as those which as it is shown in the tables above, the two corpora have certain similarities. The most obvious of them is that proportionately newspaper writing in English and Persian is dominated by referential bundles compared to the other two categories of bundles.Also, Kjellmer (1994) found that referential bundles in English newspapers are more than other bundles. As regards the other subcategories of bundles, the results indicate that there is equivalent proportionate between the numbers of bundles below these subcategories. In line withfindingsof previous studieslike Yorio (1980) English newspaper register applies topic elaboration/clarification more than the other subcategory, i.e. topic introduction/focus, this is the same in Persian journalistic writing. Regarding the "equivalent and quasi-equivalent" bundles "Cortes(2008), it is concluded that only 12 percent of bundles in English have an equivalent form in Persian corpus. Cortes (2008)defines equivalent bundles as those which.

References

Allen, D. (2009). Lexical Bundles in Learner Writing: An Analysis of Formulaic Language in the ALESS Learner Corpus.*Komaba Journal of English Education, 1*, 376.

Biber, D. S., Johansson, G., Leech, S & ,.Conrad, S .(1999) .*The Longman Grammar of Spoken and Written English*.London: Longman.

Biber, D.(2006) .*University Language: A corpus-based study of spoken and written registers* .Amsterdam: John Benjamin publishing Company.

Biber, D & ,.Barbieri, F. (2007). Lexical bundles in university spoken and written registers .*English for Specific Purpose, 26*, 286-263.

Biber, D & ,.Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard & ,S. Okesfjell (Eds) ,*Out of Corpora: Studies in Honor of Stig Johansson* (pp. 181-190). Oxford University Press.

Biber, D., Conrad, S & ,.Cortes, V. (2004). If you look at…: Lexical Bundles in university teaching and textbooks .*Applied Linguistic*.۴۰۵-۳۷۱ ,

Cortes, V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish .*Corpora*.۵۷-۴۳ ,

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology.*English for Specific Purpose, 23*, 423-397.

De. Cock, S. (2000). Repetitive phrase chunkiness and advanced EFL speech and writing .*Twentieth International Conference on English Language Research on Computerized* (pp. 51-68). Amsterdam: Rodopi.

Haswel, R(1991) .*Gaining Control in College Writing: Tables & Development & Interpretation* .Dallas: Southern Methodist University Press.

Hewings, M & ,.Hewings, A. (2002). " it is interesting to note that…": A comparative study of anticipatory "it" in student and published writing .*English for Specific Purpose, 21* ,367-383.

Hyland, K. (2008). As can be seen…: Lexical bundles and disciplinary variation .*English for Specific Purpose, 27*, 21-24

Johansson, S (2007) .*Seeing Through Multilingual Corpora: On the Use of Corpora in Contrastive Studies* .Amsterdam and Philadelphia: John Benjamins.

Kjellmer, G.,(1994) *A disciplinary of English collocations* .Oxford: Clarendon Press.

Milton, J. (1998). Exploiting L2 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger ,*learner Englishon computer* (pp. 186-198). London: Longman.

Nattinger, J. R & ,.DeCarrico, J, (1992). *Lexical phrases and language teaching* .Oxford: Oxford University Press.

Salem, A(1987) .*Pratique des Segments Répétés* .Paris: Institut National de la Langue Française.

Schmitt, N & ,.Carter, R. (2004). "Formulaic sequences in action: An introduction". In N. Schmitt ,*Formulaic sequences: Acquisition, processing and use* (pp. 1-22). London: Longman.

Scott, M. (1996). Wordsmith Tools 4. Oxford University Press.

Sinclair,J,(1991) *Corpus, Concordance, Collocation* .Oxford: Oxford University Press.

Stubbs, M & ,.Barth, I. (2003). Using recurrent phrases as text- type discriminators: A quantitative method and some findings .*Functions of Language, 10*, 104-61(1).

Yorio, C. (1980). Conventionalized language forms and the development of communicative competence .*TESOL Quarterly, 14* ,(۴)433.