Connectives in the World Wide Arabic corpus

Haslina Hassan

Nuraihan Mat Daud

International Islamic University Malaysia

Eric Atwell

University of Leeds

haslina.h@iiu.edu.my; nuraihan@iiu.edu.my; eric@comp.leeds.ac.uk

Abstract

This study analysed the use of connectives in the World Wide Arabic corpus of selected Gulf countries. The corpus was built by using Web BootCat where the Arabic sites had been extracted based on Arabic seed-words parallel to the English ones (Sharoff, 2006). A quantitative method has been employed to analyse the Arabic connectives extracted from the word lists prepared by SketchEngine. The results revealed that connectives, particularly connectives' sub-topic, namely ḥurwf al-jar "حروف الجر" (prepositions) appeared to be on the top ten list for the most frequent words used in all corpora irrespective of country and genre. The study also observed that there are few connectives listed and repeatedly cited in the Traditional Arabic Grammar but are not found in the corpus.

Keywords: Arabic corpus, Arabic connectives, second language learning

1.0 Introduction

In teaching a language one would have to choose materials that are appropriate for the level taught. As there are many aspects that need to be taught, there is a need to prioritize the elements to be focused on. These elements may differ from one language to another depending on among them the frequency of usage. Although frequency is not the only criterion for selecting what to teach, it should be given due consideration in the development and choice of materials teachers bring into classrooms (McEnery, 2006; Biber, 2002; Fox, 2001). Studies have shown that teaching words that are frequently used are more useful to students whereas rare words are less useful in the earlier stages of language learning (Biber, 2002; Fox, 2001).

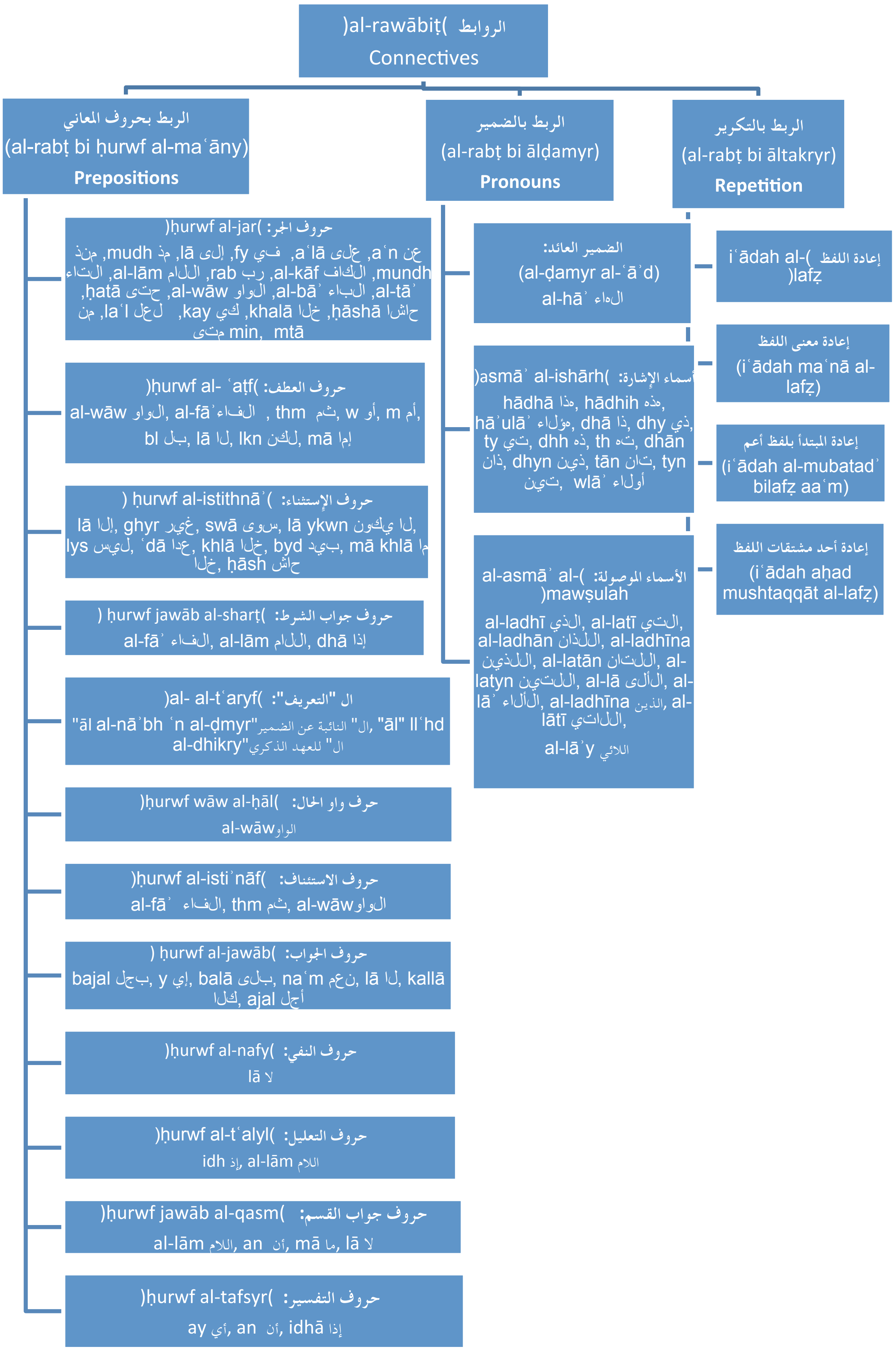
Suggestions have been made to produce materials for language instructions and assessment based on a corpus where language is presented from natural texts rather than intuition

(Biber, 2002; Fox, 2001; McCarthy, 2001; Mindt, 1996; Byrd, 1995b; McDonough & Shaw, 1993). Such an approach will expose learners particularly those in non-native environment to samples of natural language. The availability of such a corpus gives students opportunities to discover language and apply it based on the new linguistic knowledge they generate from the corpus (Hadley, 2002; Willis, 2001). This process is important as 'noticing' features of the targeted language are an inevitable stage in a learning process (Richards, 2005; Krieger, 2003; Biber & Conrad, 2001; Conrad, 1999; Schmidt, 1990). Hence, materials which are more relevant to students' needs may be produced using this approach.

Each language normally has its own unique feature. In this study, Arabic will be the basis for discussion in the use of corpus in developing teaching materials.

2.0 Features of Arabic

In Arabic, a kernel sentence is made up of *al ism* "الاسم" (noun), *al fiʿl* "الفعل" (verb), and *al ḥurwf* "الحرف" connective (Malik, cited in Ghalayaini, 1987). Connectives such as *fī* "في" (in) and *ilā* "إلى" (to) are used more frequently than nouns and verbs. The Arabic connectives signal a specific relationship. They guide the reader or listener to understand the relationship between two words or more in a sentence or what exists beyond the sentence level (Hassan, 2001). These connectives are categorized into three major components: (a) *al-rabṭ biḥurwf al-mʿāny*, (b) *al-rabṭ biālḍamyr* and (c) *al-rabṭ biāltakryr* (Hassan, 2004). The first component comprises*: ḥurwf al-jar, ḥurwf al- ʿaṭf, ḥurwf al-istithnāʾ, ḥurwf jawāb al-sharṭ, al- al-taʿryf, ḥurf wāw al-ḥāl, ḥurwf al-istiʾnāf, ḥurwf al-jawāb, ḥurwf al-nafy, ḥurwf al-taʿlyl, ḥurwf jawāb al-qasm, ḥurwf al-tafsyr.* The second is made of *al-ḍamyr al-ʿāiʾd , asmāʾ al-ishārah* and *al-asmāʾ al-mawṣulah* . Lastly *al-rabṭ bi altakryr* which include *iʿādah al-lafẓ ,iʿādah maʿnā al-lafẓ ,iʿādah al-mubtadaʾ bilafẓ aaʿm* and *iʿādah aḥad mushtaqqāt al-lafẓ*. (Please refer to Table 1 for the details).



)al-rawābiṭ(الروابط
Connectives

الربط بحروف المعاني
(al-rabṭ bi ḥurwf al-maʿāny)
Prepositions

)ḥurwf al-jar(:حروف الجر
عن a‘n, على a‘lā, في fy, إلى lā, مذ mudh, منذ mudh, الكاف al-kāf, رب rab, اللام al-lām, التاء al-tāʾ, الباء al-bāʾ, الواو al-wāw, حتى ḥatā, حاشا ḥāshā, خلا khalā, كي kay, لعل laʿl, من min, متى mtā

)ḥurwf al- ʿaṭf(:حروف العطف
al-wāw الواو, al-fāʾالفاء , thm ثم, w أو, m أم,
bl بل, lā لا, lkn لكن, mā إما

) ḥurwf al-istithnāʾ(:حروف الإستثناء
lā إلا, ghyr غير, swā سوى, lā ykwn لا يكون, lys ليس, ʿdā عدا, khlā خلا, byd بيد, mā khlā ما خلا, ḥāsh حاش

) ḥurwf jawāb al-sharṭ(:حروف جواب الشرط
al-fāʾ الفاء, al-lām اللام, dhā إذا

)al- al-tʿaryf(:"التعريف" ال
"āl al-nāʾbh ʿn al-ḍmyr"النائبة عن الضمير "ال, "āl" llʿhd al-dhikry"للعهد الذكري "ال

)ḥurwf wāw al-ḥāl(:حرف واو الحال
al-wāwالواو

)ḥurwf al-istiʾnāf(:حروف الاستئناف
al-fāʾ الفاء, thm ثم, al-wāwالواو

) ḥurwf al-jawāb(:حروف الجواب
bajal بجل, y إي, balā بلى, naʿm نعم, lā لا, kallā كلا, ajal أجل

)ḥurwf al-nafy(:حروف النفي
lā لا

)ḥurwf al-tʿalyl(:حروف التعليل
idh إذ, al-lām اللام

)ḥurwf jawāb al-qasm(:حروف جواب القسم
al-lām اللام, an أن, mā ما, lā لا

)ḥurwf al-tafsyr(:حروف التفسير
ay أي, an أن, idhā إذا

الربط بالضمير
(al-rabṭ bi ālḍamyr)
Pronouns

:الضمير العائد
(al-ḍamyr al-ʿāʾd)
al-hāʾ الهاء

)asmāʾ al-ishārh(:أسماء الإشارة
hādhā هذا, hādhih هذه, hāʾulāʾ هؤلاء, dhā ذا, dhy ذي, ty تي, dhh ذه, th ته, dhān ذان, dhyn ذين, tān تان, tyn تين, wlāʾ أولاء

al-asmāʾ al-(:الأسماء الموصولة
)mawṣulah
al-ladhī الذي, al-latī التي, al-ladhān اللذان, al-ladhīna اللذين, al-latān اللتان, al-latyn اللتين, al-lā الألى, al-lāʾ الألاء, al-ladhīna الذين, al-lātī اللاتي,
al-lāʾy اللائي

الربط بالتكرير
(al-rabṭ bi āltakryr)
Repetition

iʿādah al-(إعادة اللفظ
)lafẓ

إعادة معنى اللفظ
(iʿādah maʿnā al-lafẓ)

إعادة المبتدأ بلفظ أعم
(iʿādah al-mubatadʾ bilafẓ aaʿm)

إعادة أحد مشتقات اللفظ
(iʿādah aḥad mushtaqqāt al-lafẓ)

Table 1: Connectives in Arabic

The main difference with English is that in Arabic a complete sentence can consist of only connectives, or only a connective and a noun, or two or more connectives and a noun.

Example of a complete sentence with connectives only:

- "في غير ه" (fy ghayri hi)

  (it) (except) (in)

  In the other's.

  The word *fī* "في" above represents what is termed as *ḥrwf al-jar* in Arabic which is similar to preposition in English. The word *ghayr* "غير" is categorized as *ḥrwf al-istithnāʾ* and *hi "ه"*is under *al-ḍamyr*. In Arabic all these three are considered as connectives.

Example of a complete sentence with a connective and a noun:

- "مع السلامة" (maʿa al-salāmah)

  (peace) (with)

  Bye.

  The word *maʿa* "مع" above is termed as *ḥarwf al- ʿaṭf* in Arabic and categorized as a connective.

Example of a complete sentence with two connectives and a noun:

- "لـه حق" " lahu ḥaq"

  (right) (he) (for)

  He has the right.

  The sentence above is made up of two connectives: *ḥarwf al-jar* "ل" , *al-ḍmyr* "ه" and a noun *ḥaq* " حق".

Example of a complete sentence with two or more connectives and a noun:

- "هذا الذي قلته" (hādhā al-ladhī qultuhu)

  (it) (I said) (which) (this)

This is what I said.

The word *hādhā* "هذا" is categorized as *ism al-ishārah, al-ladhī* "الذي" is under *ism al-mawṣwul* and *hū* "ه" as *al-ḍamyr*. All these three are considered as connectives.

In any language, the wide range of connectives and the multiple-meaning each carries in a particular context of utterance makes the teaching of connectives challenging and difficult for the learners to put them to use (Tapper, 2005; Fox, 2001; Granger & Tyson, 1996; Wikborg & Bjork, 1989). In a non-native environment, students may be at a disadvantage because they may not be exposed to all the contexts of occurrence for the various Arabic connectives. Barlow (2002, cited in Krieger, 2003) suggests that one way of solving this problem is by using a corpus in materials development. However, this kind of corpus is not easily available in the Arabic world. This does not mean that the approach cannot be applied on Arabic language teaching. The use of an appropriate concordancer may allow the adoption of such an approach since in digital format are easily available on the Internet. This study will look into the possibility of using a concordancer to study the connectives that are frequently used in selected Arabic speaking countries. It will focus on the first two components only: *al-rabṭ bi ḥarwf al-maʿāny* and *al-rabṭ bi ālḍamyr,* as these two comprise specific fixed words and can be easily identified in a text. Whereas the last component which is *al-rabṭ bi āltakryr* is made up of lexical items that vary from one context to another.

3.0 Statement of the Problem

Although there are fifteen different types of connectives in Arabic, not all of them are used frequently by its speakers. Hence there is a need to identify the frequency of use for each group to help in identifying which type should be focused on and what to be taught first in teaching Arabic particularly to foreign learners. This study is thus conducted to find the frequency of connectives usage by the native speakers of Arabic.

4.0 Objectives of Study

The objectives of this study are to:

(i) find the number of times a connective occurred in selected Arab countries websites, particularly those that are based in Saudi Arabia, Egypt, Jordan, Sudan and Iraq;

(ii) see whether there are differences in the frequency of use of connectives in these five Arab countries.

5.0 Methodology

Data in this study was drawn from Internet materials from five of the main Arabic speaking countries namely Saudi Arabia, Egypt, Jordan, Sudan and Iraq. The data that was compiled using WebBootCat and Sketch Engine was then applied to search by keyword-in-context. A corpus which consists of about 200,000 words from each WWW national domain was developed. This was done by restricting the search to the country sites based on the country domain such as URLs which end with ".eg" for Egypt, ".sa" for Saudi Arabia, ".jo" for Jordan, ".sd" for Sudan, and ".iq" for Iraq.

WebBootCat was used to find the lists of URLs which match subsets of 3 seed-words, and to generate webpages listing the URLs. The seed-words refer to common words that appear in any ordinary language text; either single word or multi-word expressions (Baroni, 2006). This study used Arabic seed-words prepared by Latifah Al Sulaiti (2006) which is parallel to the English ones prepared by Serge Sharoff (2006).

The size of data was, however, limited by the capacity of the software. At one time, the software could analyse up to 1 million tokens only. For the purpose of this study, an equal number of text size which is around 200,000 words from each domain was analysed. The total size of the corpus compiled was 1,002,042 words.

6.0 Analysis of Results

Of the 1,002,042 word corpus of Arabic, the *ḥrwf al-jr* (prepositions) were found to be the most frequently used connectives (see Table 2 for the list of number of occurences).

DOMAIN/ SIZE(word)	EGYPT (200,413)			IRAQ (200,225)			JORDAN (200,792)			SAUDI ARABIA (200,091)			SUDAN (200,521)		
RANKING	WORD	FREQ	%	WORD	FREQ	%	WORD	FREQ	%	WORD	FREQ	%	WORD	FREQ	%
1	في	2155	1.08%	في	2091	1.04%	في	2474	1.23%	في	4629	2.31%	في	2212	1.10%
2	من	2005	1.00%	من	1795	0.90%	من	1894	0.94%	من	3767	1.88%	من	1584	0.79%
3	على	951	0.47%	على	1045	0.52%	أو	1162	0.58%	على	1945	0.97%	على	865	0.43%
4	أن	919	0.46%	و	905	0.45%	على	1147	0.57%	أن	1194	0.60%	أن	519	0.26%
5	إلى	651	0.32%	أن	543	0.27%	أن	880	0.44%	إلى	851	0.43%	ان	452	0.23%
6	بعد	197	0.10%	عن	538	0.27%	إلى	631	0.31%	عن	783	0.39%	التي	378	0.19%
7	ولا	184	0.09%	الله	479	0.24%	التي	534	0.27%	الله	767	0.38%	عن	367	0.18%
8	ذلك	183	0.09%	ما	439	0.22%	و	430	0.21%	أو	729	0.36%	ما	316	0.16%
9	ب	177	0.09%	لا	433	0.22%	م	424	0.21%	التي	693	0.35%	إلى	309	0.15%
10	الحوار	167	0.08%	إلى	422	0.21%	عن	420	0.21%	لا	617	0.31%	الذي	294	0.15%
11	كانت	154	0.08%	ان	372	0.19%	هذه	414	0.21%	هذا	570	0.28%	مع	292	0.15%
12	عليه	153	0.08%	هذا	361	0.18%	ما	405	0.20%	و	512	0.26%	الى	281	0.14%
13	عام	152	0.08%	كل	304	0.15%	لا	381	0.19%	مع	502	0.25%	لا	272	0.14%
14	قبل	151	0.08%	التي	287	0.14%	هذا	344	0.17%	ما	482	0.24%	الله	272	0.14%
15	حتى	151	0.08%	أو	287	0.14%	ي	327	0.16%	الذي	434	0.22%	هذا	252	0.13%
16	ك	148	0.07%	هذه	275	0.14%	ذلك	271	0.13%	هذه	433	0.22%	هذه	247	0.12%
17	إن	147	0.07%	فى	274	0.14%	كان	248	0.12%	م	383	0.19%	هو	244	0.12%
18	ه	144	0.07%	هو	271	0.14%	أي	237	0.12%	ذلك	375	0.19%	و	211	0.11%
19	أي	133	0.07%	مع	232	0.12%	ل	210	0.10%	بين	367	0.18%	أو	195	0.10%
20	قد	132	0.07%	كان	223	0.11%	ص	203	0.10%	ان	358	0.18%	كان	191	0.10%

Table 2: Occurrence of Connectives *ḥarwf al-jar* ▨

In all the five countries, *ḥarwf al-jar* was highly employed in the selected texts. This is followed by *ḥarwf al- aʿṭf,* then *al-asmāʾ al-mawṣwulah* and *asmāʾ al-ishārah.* Such an information may be used by teachers in deciding which connectives is to be taught first to Arabic learners. This finding is in line with suggestions made by Biber (2002) and Mindt (1996) that the order of grammatical topics should be based on frequency study. In this case, it is advisable to teach *ḥarwf al-jar* to the beginners followed by *ḥarwf al-a ʿṭf* , then a*smāʾ al-ishārah* and *al-asmāʾ al-mawṣwulah*. *Ḥarwf al-istiʾnāf* and *ḥarwf al-tafsyr* may be stressed on in the advanced level classes.

The table also shows that the connective word has multiple meanings with word with a certain connotation occurring more frequently than others. For example the word `min` "من":

(1) " مشيت من المكتبة إلى المسجد" "mashaytu min al-maktabah ilā al-masjid"

(mosque) (to) (library) (from) (I) (walked)

I walked from the library to the mosque.

The *ḥarwf al-jar* “من” in sentence (1) signifies *ibtidāʾ al-ghāyah* (starting point).

(2) " أكلت جزءاً من الرغيف" "aʾkaltu juzuʾʾ min al-raghyf"

(bread) (from) (part) (I) (ate)

I ate part of the bread.

The *ḥarwf al-jar* “من” in sentence (2) means *al-tabiʿyḍ* (part of).

(3) " قربت منه" "qarabtu minhua"

(him) (from) (I) (closed)

I came close to him.

The *ḥarwf al-jar* “من” in sentence (3) indicates *al-intihāʾ* (ending).

(4) "المدير يعرف الطالب المجتهد من الطالب المتكاسل"

"ālmudyr yaʿrf al-ṭālib al-mujtahid min al-ṭālib al-mutakāsil"

(lazy) (student) (from) (hardwork) (student) (knows) (headmaster)

The headmaster can distinguish a hardworking student from a lazy one.

The *ḥarwf al-jar* “من” in sentence (4) means al- fa□l (distinguish)

Hence a syllabus designer would also need to consider deciding which meaning of the same word should be stressed on first in teaching the language.

The existence of the corpus itself may help the teacher or material developer in providing examples of sentences based on their context of occurrence. The raw data can be a rich source for material development. An example for *ḥarwf al-jar* that can be extracted from the corpus is:

وهكذا تسير الدولة منذ نشأتها في خطوات ثابتة نحو توثيق عرى التوحد و الترابط

(SketchEngine, doc.id 2 , doc.text 715-6296)
Since its inception the state moves steadily towards a closer unity and coherence.

Table 2 also shows that the frequency of occurrence of the different types of connectives is the same in all the countries chosen for this study. This reflects that there is a specific pattern of usage in the real world.

The analysis also revealed that some of a*smā ʾ al-ishārah* which belonged to *al-rabṭ bi ḥarwf al-ma ʿāny* did not appear in the frequency list. These are, however, included in many Arabic language textbooks including those meant for beginners e.g. *tānikum* تانكم, *tānikun*تانكن , *dhānikum* ذانكم , *ulāʾikuma* أولئكما , *tānika* تانك , *hātyna* هاتين , *dhākunn* ذاكن , *al-alāʾ* الألاء , *dhānikun* ذانكن. Often students are expected to memorize their usage although they hardly encounter these words in their daily life.

7.0 Conclusion

The study shows that there is an order in the frequency of usage of the Arabic connectives. Such an order is not only limited to the specific group but it is observed that certain meaning of a particular word occurs more frequently than its other connotations. Such information is valuable in deciding materials to be taught to learners.

It is high time that language materials development in general and the designer of the Arabic grammar syllabus in particular is informed by data sourced from a corpus, as this provides authentic language use and facilitate language learning. Such a move will make teaching more relevant and useful to the learners of the language. Future efforts should be focused on how to convince the Arabic teachers that corpus based materials can promote discovery learning in the classroom.

Data-driven approach should be popularized. Teachers and researchers alike should make use of web-tools like BootCat and SketchEngine that make it possible for them to collect their own corpus and do data-driven much more easily, without having to be technical experts in programming, so long as they understand basic web-as-corpus ideas like defining a seed-term list of words typical of the language they are teaching and investigating.

References:

Barlow, M. (2002). *Corpus Linguistics: What It Is and How It Can Be Applied to Teaching.* Retrieved November 7, 2009, from iteslj.org/Articles/Krieger-Corpus.html

Baroni, M. B. (2006). *WebBootCaT: instant domain-specific corpora to support human translators.* Retrieved November 2009 7, 2009, from http://trac.sketchengine.co.uk/wiki

Biber, R. R. (2002). What Does Frequency Have to Do with Grammar Teaching. *Studies in Second Language Acquisation (24)* , 199-207.

Biber, S. C. (2001). Quantitave Corpus-Based Research: Much More Than Bean Counting. *Teachers of English to Speakers of Other Languages (TESOL) Quarterly* , 331-336.

Cermakova, W. T. (c2007). *Corpus Linguistics : A Short Introduction.* London and New York: Continuum.

Conrad, S. (1999). The Importance of corpus-based reasearch for language teachers. *System* , 1-18.

Fox, G. (2001). Using corpus data in the classroom. In B. Tomlison, *Materials Development in Language Teaching* (pp. 25-43). Cambridge: Cambridge University Press.

Gahlayaini, M. A. (1987). *Jām' Al-Drws Al-'Rbyt.* Beirut: Al Maktabah al Asriyyah.

Hadley, G. (2002). An Introduction to Data-Driven Learning. *RELC Journal 33(2)* , 99-124.

Halliday, M. (2004). *Lexicology and corpus linguistics : an introduction* . London ; New York: Continuum.

Hassan, T. (2001). *Al Lughah Al 'Arabiyyah: Ma'naha Wa Mabnaha.* Al Dar Al Baida': Dar Al Shaqafah.

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge : Cambridge University Press.

K., B., G., N., & J., H. (2003). *A corpus-based study of connectors in student writing: Research from the International Corpus of English in Hong Kong (ICE-HK)* . Retrieved March 8, 2010, from International Journal of corpus linguistics: http://www.ingentaconnect.com/content/jbp/ijcl/2003/00000007/00000002/art00002

Krieger, D. (2003, March). Retrieved March 10, 2010, from http://iteslj.org/

Krieger, D. (2003, March). *Corpus Linguistics: What It Is and How It Can Be Applied to Teaching.* Retrieved November 8, 2009, from The Internet TESL Journal, Vol. IX, No. 3, March 2003: http://iteslj.org/Articles/Krieger-Corpus.html

Krishnamurthy, W. T. (2007). *Corpus linguistics : critical concepts in linguistics.* London: Routledge .

Leech, G. (1997). Teaching and Language Corpora. In S. F. Anne Wichmann (Ed.), *Teaching and Language Corpora* (p. 343). United States of America: Longman.

McCarthy, R. C. (2001). Size Isn't Everything: Spoken English, Corpus, and the Classroom. *Teachers of English to Speakers of Other Languages (TESOL)* , 337-340.

Mindt, D. (1996). English corpus linguistics an dthe foreign language teaching syllabus. In M. S. Thomas, *Using Corpora forLlanguage Research* (pp. 232-247). New York: Longman Group Limited.

Richard, J. C. (2005). Materials Develpoment and Research.

Sinclair, J. (1991). *Corpus, concordance, collocation* . Oxford : Oxford University Press.

Sinclair, J. (2004). *Developing Linguistic Corpora:a Guide to Good Practice*. Retrieved May 4, 2009, from http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm

Tapper, M. (2005). *Connectives in advanced Swedish EFL learners' written English -preliminary results.* Retrieved March 9, 2010, from www.sol.lu.se/engelska/dokument/wp/vol05/Tapper-wp-05.pdf

Tony McEnery, R. X. (2006). *Corpus-Based Language Studies.* New York: Routledge.

Willis, J. (2001). Concordances in the classroom without a computer: assembling and exploiting concordances of comon words. In B. Tomlinson, *Materials Development in Language Teaching* (pp. 44-66). Cambridge: Cambridge University Press.