حجم البيانات في العينات ثنائية اللغة في دراسات التحليل الصنفي التقابلي بين العربية والإنكليزية

أمجد كاظم عبيد الركابي
وزارة التربية/مديرية التربية في بابل

Data Size for Bilingual Comparable English and Arabic Corpus in Contrastive Genre Analysis

Amjed Kadhim Al-Rickaby
Ministry of Education/Directorate of Education in Babel
Amjed.linguist@gmail.com

الملخص

يتناول هذا البحث آلية إيجاد عينة ملائمة للدراسات التقابلية بين اللغة العربية واللغة الإنكليزية. تفترض الدراسة ان معايير إيجاد عينة للدراسات التقابلية بين الإنكليزية واللغات الأوروبية الأخرى لا يمكن اعتمادها عند اجراء دراسات تقابلية بين الإنكليزية واللغة العربية ذلك ان أهم معايير إيجاد العينة وهو الحجم يعتمد على عدد الكلمات. ولاختبار صحة الفرضية، تمت دراسة عينتين تتألفان من خمس مقالات افتتاحية إنكليزية وخمس مقالات افتتاحية عربية وحُللتا من حيث مؤشرات الموقفية فيهما. ومن خلال دراسة الخصائص الصرفية للكلمات لوحظ فرق واضح في نسبة مؤشرات الموقفية. تستنتج الدراسة ان إيجاد عينة مناسبة يجب ان يكون من خلال أخذ الخصائص الصرفية للغتين بعين الاعتبار.

الكلمات المفتاحية: اللغات التوليفية، اللغات التحليلية، العينة، تحليل الصنف

Abstract

The main objective of this study is to ruminate building an equal corpus for English and Arabic contrastive analyses. The essential hypothesis is that the models devised to contrast corpora across English and other European languages might not ensure a parallel accuracy when a synthetic language is contained. This hypothesis is centered on one of the main parameters of building corpora, namely data size. In contrastive studies, data size is based on counting the number of words in the corpus. Arabic is morphologically different from analytic languages. Therefore, word counts cannot be taken at face value. To verify the hypothesis, a sample of five English editorials and five Arabic editorials were analyzed for stance markers. Taking the morphological nature of synthesized words, the study shows a significant change in frequency percentages.

Key words: synthetic languages, analytic languages, genre analysis

1. Introduction

Establishing a strict criterion for deciding on the corpus is a key factor in maintaining accurate and authentic corpus linguistic studies. Many studies have dealt with the challenges on this issue and models have been put for the same end. They, in fact, worked well when the languages involved share maximum level of similar characteristics.

However, when these models are applied to compare/contrast languages with different characteristics, accuracy can no longer be guaranteed. Building a comparable corpus is based on a number of parameters. One of these is corpus size. Talking about corpora size in linguistic studies straightforwardly means the number of words in the corpora. Many studies look at the frequency or occurrence of particular linguistic features (e.g. hedges) in a particular language domain, and hence draw conclusions basing on the frequency of these linguistic features against the total number of words in the corpora. Taking the note into a narrower and deeper level and focusing on the English versus Arabic contrastive studies as is the aim of this study, equality in word number does not always mean equality per se and the reason behind this is that English is an analytic language whereas Arabic is a synthetic one. In analytic languages, words are not combined together while in synthetic languages many words are combined together into one-word form. One single word form from Arabic may translate into a full English sentence of Eight words. If we look at the Arabic word (فَأَسْقَيْنَاكُمُوه) (then We give it to you to drink), we can see that it is made up of eight words.

Stemming from this observation, the current study attempts to point out the shortcomings of ascertaining a tertium comparationis for English and Arabic contrastive studies basing on mere words counting. It also endeavors to provide a solution for this shortcoming.

2. Words frequencies and text analysis

Recently, there has been wide interest in the text features that signal the language domain from which the data is taken (e.g. academic, professional, journalistic, etc.) and the endeavor is even taken into a further end when text features are looked for in at least two languages from the same domain to reveal cultural and community norms. Studies of this kind include, among all, metadiscourse studies, stance and engagement markers, interactive resources, discourse markers, rhetorical features, etc. Conducting these studies cannot be done without knowing the exact number of words contained in the corpus so as to count the linguistic features searched against the whole number of words. Counting these against the total number of words reveal to what extent a certain linguistic feature is common or rare and which among them is associated with a certain genre or register. This is true in intra-lingual and inter-lingual studies equally. In Hyland's study (2005) of stance and engagement markers, he used a monolingual corpus of 1.4 million words and "searched for specific features seen as initiating writer–reader interactions" (Hyland, 2005: 178). The corpus of his study is taken from the English language. The corpus findings show that stance and engagement markers occur about one every 28 words with occurrences of 200 in each paper. The significance of these frequencies would not have been available if the total number of words tokens had not been known.

On a similar vein, Sultan (2011) conducted a contrastive study to examine the implementation of metadiscoursal resources in English and Arabic research article abstracts. For this contrastive study, the total amount of the English corpus is 23,903 words and the Arabic corpus is 25,552 words. The 70 articles are collected from international academic journals written by English and Arab linguists. The researcher used a Chi square in a null hypothesis to test the differences between the English and the Arabic corpus and he found that the value of observed chi-square ($x2 = 15.97$) is meaningful at α level ($\alpha = 0.05$) with a freedom degree of 4. This result denotes a noteworthy difference in the employment of interactive metadiscoursal resources between Arabic and English writers. And for the interactional resources, the researcher found that the observed value of chi-square ($x2 = 13.10$) is significant at α level ($\alpha = 0.05$) with a freedom degree of 4. The examination of the total corpus reveals that there are 2,296 metadiscoursal markers in the 49,455 words corpora. Put in plain words, there is one metadiscoursal marker in every 21 words. Split into Arabic and English corpora, this is one metadiscourse marker per 23 words for the English corpus (total English corpus 23,903 words), and one metadiscourse marker per 20 words for the Arabic corpus (total Arabic corpus 25,552 words).

3. Corpus design and contrastive studies

With the advancement of computational tools and accessibility to huge amounts of texts, massive steps forward has been taken in the ripeness of relatively new discipline in linguistics called corpus linguistics. Corpus linguistics has become a base stone in many linguistic studies (especially, translation studies, discourse studies, text analysis, contrastive studies) due to the authenticity it gives to the research findings when used aptly. This vast development in the new field requires developing new methods of application for an upper limit of utility.

In accordance with the aim and nature of study, different types of corpora have been founded. Characteristically, two main types are defined; parallel corpora (bi- or multi-lingual) which is composed of source and target texts, and comparable corpora, defined as corpora created according to similar design criterion (Fantinuoli and Zanettin, 2015: 3). A more detailed taxonomy is introduced by Mcenery and Xiao (2010: 2) where two criteria have been adopted in defining the different types of corpora; the number of languages involved and the form of the content. By number of languages involved corpora types are classified into monolingual, bilingual, and multilingual, whereas by form, they are classified into parallel corpus (source text and their

translations in parallel) and comparable corpus (matched samples from different languages). A bilingual comparable corpus can be broadly defined as a corpus containing two sets of data collected using identical sampling frame and equal balance and representativeness. This sampling method entails "the *same proportions* of the texts of the *same genres* in the *same domains* in a range of *different languages* in the *same sampling period*" (ibid). However, the two sets of data in a bilingual comparable corpus are not translations of one language into another. Rather, their comparability comes from the sampling method and the equal quantities of input. Our focus here is on bilingual comparable corpus. Typically, building a comparable corpus is governed by a number of parameters to guarantee comparing real comparable corpora. These parameters are discussed by Moreno (2008).

4. Analytic and Synthetic Languages and contrastive linguistics

A genre can be studied comparatively or contrastively across two or more languages and/or cultures basing on a platform of sameness (tertium comparationis). This platform includes deploying a matching method that ensures equal quantities and representativeness. This sampling method can be summarized as follows:

a. Same proportions of the texts (equal data size)
b. Same genres
c. Same domains
d. Same sampling period
e. Different languages

Therefore, to establish a platform of sameness, the above-mentioned concerns must all be taken into consideration. This process of building a bilingual comparable corpus is more complicated when the languages involved are of different characteristics. The current paper tackles the concern of founding equal data size as a parameter of ensuring sameness in English and Arabic bilingual comparable corpus. In contrastive studies, data size is of key importance in rendering the findings reliable and truly reflective. Because of the fact that data size is always calculated on the basis of the number of words contained, the exactness of establishing equal data size cannot be easily guaranteed when English and Arabic are compared due to the morphological characteristics of their words. Arabic words, for instance, are often composed of two parts of speech, e.g. noun + pronoun, (حلم *his dream*) subject + verb (حلمت *I dreamt*).

To elaborate further on this, reference should be made back to the terms analytic languages and synthetic languages which are embraced in cross-linguistic typology using structural principles and centering on the words characteristics. In analytic languages, on the one hand, words are unwavering and syntactic relations are mainly exhibited by word-order. Synthetic languages, (which also cover inflecting and agglutinative types), on the other hand, typically allow their words to contain more than one part of speech. As usually, categories under such classifications are not clear-cut: some languages may display the features of analyticity to a lessr or greater extent (Crystal, 2008: 24-5). Put precisely, the distinction between synthetic and ananltyic languages is not a "bipartition or a tripartition, but a continuum, ranging from the most radically isolating to the most highly synthetic languages" (Haspelmath and Sims, 2010: 5). We can decide the position of a certain language on this continuum by exploring its extent of analyticity or synthesis, i.e. the number of morphemes or parts of speech per word in a representative random text sample of that language (ibid).

Additionally, Richards and Schimdt (2002: 31) point out that word forms do not change in analytical languages and function words and word order show the grammatical relations. Some languages such as Chinese and Vietnamese are considered highly isolating analytical languages. Researchers maintain that there is no straightforward distincting tool between inflecting languages, isolating languages, and agglutinating languages. In the family of European languages, English is most isolating one among French, German, and Russian, yet it is also termed an inflecting language. In contrast, various affixes can be attached to the stem of a word in synthetic languages to change

its meaning or to display a certain grammatical function. For example, the word انقذونا *anqathuna* (they saved us) is a full sentence in Arabic, composed of a subject pronoun, a past form of the verb and an object pronoun. Other examples of synthetic languages include Finnish, Hungarian, Swahili, and Turkish.

5. Corpus linguistics

One of the fastest mounting fields in present-day linguistics is corpus linguistics. It can precisely be described as a turn in the way researchers locate and utilize data (Joseph, 2004: 382). As a methodology, descriptive and applied ends are the main focus of large part of corpus linguistic research. They are always grounded on the investigation of some types of occurrences and frequencies (Gries, 2009: 1-2). More expressly, this research is concerned with whether:

1. something exists in a certain corpus; i.e. whether the observed frequency (of occurrence or co-occurrence) is zero or larger;
2. something exists in a certain corpus more frequently than something else; i.e. whether an observed frequency is larger than the observed frequency of something else;
3. something exists more or less frequently than we would expect by chance (ibid, 3)

Examples of linguistic research where corpus linguistics plays a prime role are so many. Among these is cross cultural and cross linguistic studies which have been strengthened by the authenticity given by corpus linguistics. Parington (2004: 44) points out that cross-cultural studies come under criticism for a lack of systematicity in the past. Now, the kind of comparative statistical analyses which corpus techniques makes available constitute an extremely valuable way of providing quantative evidence regarding similarities and differences across languages and cultures.

A corpus (plural corpora) can be "anything from a small set of texts parsed and tagged right up to multi-million word digitized collections of spoken and written data" (Trask, 2007:49). With the advancement of technology, interest has shifted to electronic corpora (or computer corpora). The term corpus linguistics is now generally associated with the utilization of such corpora. Corpora have been built and exploited for research on English and many other languages. The term corpus linguistics which has come into existence since the early 1980s is mostly used in linguistic research which depends on the use of computer (Malmkjar, 2002: 104-5).

6. Data collection and method of analysis

The concept of *tertium comparationis* or common ground of comparison is of prime importance at all levels of the research in contrastive linguistics: in classifying texts for establishing comparable corpus, choosing textual features to be attested in the corpora, and recognizing and distinguishing between different kinds of linguistic resources implemented to realize these features. To comply with all these concerns, Connor and Moreno (2005) projected a model for contrastive studies (see the table below). This model starts with establishing comparable corpora basing on pertinent similarity constraints (e.g. the text form, genre, the level of writers' expertise, mode of communication, etc.) that may have effect(s) on the expression of the textual concept(s) in question. Then, their model suggests to build common platform not only on the text conceptual or functional level but also on the level of text realization for an equivalent match of any two sets of data before successful quantitative comparisons can be performed.

Figure (1): corpus building similarity constraints

Tertium comparationis	
Text form	Argumentative texts
Genre	Newspaper editorial
Mode	Written language
Participants	
• Writers	Editorialists

• Targeted readers	Average people
Situational variety	Formal
Dialectal variety	Standard
Tone	Serious
Channel	Paper and electronic material
Formal features	
• Length • Intertextuality • Visual features	Reference to other texts None
Point of view	Subjective
Global communicative event	Sharing opinions and influencing other's opinions and actions
Setting	Home, workplace, etc.
General purpose of communication	Writer's viewpoint: To persuade the readers to share the writer's viewpoint Reader's viewpoint: To improve one's knowledge about a given event or case
Global rhetorical strategy	Demonstrating the writer's viewpoint
Overall subject-matter or topic	Politics
Level of expertise	Professional writers
Textual unit of analysis	Complete texts
Global superstructure	(1) Presenting the case (2) Offering the argument (3) Reaching the verdict (4) Recommending action
Predominant text-types	Argumentation

Adopted from Connor and Moreno (2005) model

7. Data collection

The aim of this paper is not to investigate a certain linguistic feature but to test out to what extent the parameter of data size is dependable as a similarity constraint for establishing bilingual comparable corpus across English and Arabic contrastive genre studies. Five newspaper editorials were taken from each language. The English editorials are taken from the Independent Newspaper

and the Arabic editorials are taken from AL-Quds Al-Arabi Newspaper. Both newspapers are known as daily political independent quality newspapers and the editorials selected cover the same event. These were published during what is called "the Arabic Spring" and covered the Syrian and Libyan crises

8. Methodology

The method of analyzing the data is composed of basic three steps. In the first step, the data from both English and Arabic corpora were searched for stance markers following Hyland (2005) model of stance and engagement markers. The results will be displayed in a table and comments on the frequency of these markers against the total number of words will be made. Then in the second step, the synthesized words in the Arabic corpus will be analyzed into their constituent parts. For example, the synthesized word (باستقباله *in receiving him*) will be analyzed into (بـ ، استقبال – ـه). In the same step, the words will be counted again after the analysis. Finally, in the third step, the frequency of the same stance markers will be calculated against the new total number of words and comparison of the frequency of occurrence of these markers will be made across the two corpora.

9. Data analysis and findings

These two sets of data were searched for stance markers only (the first part of Hyland's (2005) model of stance and engagement markers) because the aim of the study here is not about how these markers are used but to see whether the percentages of their frequencies differ when synthesized words are taken into consideration. The English corpus which is composed of (2162) words included (101) stance markers. The Arabic corpus, on the other hand, is composed of (2383) words and included (114) stance markers. Detailed statistics of the subcategories of these markers are shown in the table below for both English and Arabic corpora.

Table (1): Numbers and percentages of stance markers in English and Arabic corpora

Subcategories	English		Arabic	
	Total number	Percentage	Total number	Percentage
Hedges	34	1.57	17	0.71
Boosters	37	1.71	63	2.64
Attitude markers	26	1.20	20	0.83
Self-mention	4	0.18	14	0.58
Total	101	4.67	114	4.78

After analyzing the synthesized words into their constituent parts, the total number of words in the Arabic corpus jumped to (3503) words and the percentage of these stance markers against the total number of words in the Arabic corpus decreased to (3.25%). The new statistics are presented in table (2) below.

Table (2): new percentages of stance markers in the English and Arabic corpora

Subcategories	English		Arabic	
	Total number	Percentage	Total number	Percentage
Hedges	34	1.57%	17	0.48%
Boosters	37	1.71%	63	1.79%
Attitude markers	26	1.20%	20	0.57%
Self-mention	4	0.18%	14	0.39%
Total	101	4.67%	114	3.25%

The analysis above reveals that the difference in words number is meaningful and noteworthy and should not be ignored when building a corpus involving English and Arabic data. The change in the percentages of these markers changed notably in this small sample after the synthesized words had been analyzed into their constituent parts. This change is definitely of key importance when the study objective is counting the occurrence frequency of certain linguistic features (stance markers frequency in this study dropped from 4.78% to 3.25 in the same corpus).

Recent English and Arabic contrastive studies focusing on linguistic features frequencies that reveal cultural implications have to deal with this issue very carefully. To investigate these linguistic features frequencies, the number of the words in the corpus and frequency of the features in question play a significant role. For instance, to explore the use of boosters and to count for the cultural inferences of their implementation, the study has to take into consideration the frequency of the linguistic forms that function as boosters and calculate it against the total number of words in the corpus.

10. Conclusion

The base stone for contrastive studies focusing on linguistic features that reveal cultural implications basing on words occurrence and frequency is well-balanced corpus that ensures maximum level of similarity constraints. The value of words occurrence and frequency cannot be determined without knowing the exact number of words in the sample. This point in particular is discussed and investigated in this study. Establishing a bilingual comparable corpus for a contrastive English (as an analytic language) and Arabic (as a synthetic language) study necessitates a careful consideration of the corpus size. Because corpus size is normally measured by words number, equality cannot be guaranteed very easily due to the nature of the morphological structure of words in Arabic. Arabic is a synthetic language and this fact means that a word in Arabic can be composed of more than one part of speech. Each part of these can be a word in English. It can be a pronoun, preposition, definite article. This, of course, affects the accuracy of the gained results and consequently the value of the research findings because inaccurate percentages of occurrence and frequency do not reveal true cultural norms.

References

Connor, Ulla M. & Moreno, Ana I. (2005). Tertium Comparationis: A vital component in contrastive research methodology. In P. Bruthiaux, D. Atkinson, W. G. Eggington, W. Grabe, & V. Ramanathan (eds), Directions in Applied Linguistics: Essays in Honor of Robert B. Kaplan. Clevedon, pp. 153-164. England: Multilingual Matters.

Crystal, D. (2008). A Dictionary of Linguistics and Phonetics. New York: Blackwell Publishing.

Fantinuoli, C and F. Zanettin (2015). Creating and using multilingual corpora in translation studies. In C. Fantinuoli and F. Zanettin (ed.) New directions in corpus-based translation studies. Berlin: Language Science Press.

Gries, S. T. (2009). What is Corpus Linguistics? Language and Linguistics Compass. 3(5): 1225-1241.

Haspelmath, M., and Sims, A. D. (2010). Understanding Morphology. London: Hodder Education.

Hyland, K. (2005). Stance and engagement: a model of interaction in academic discourse. Discourse Studies 7(2): 173–192.

Joseph, B. (2004). On Change in Language. Language. 80.381-4.

Malmkjar, k. (2002). The Linguistics Encyclopedia. New York. Routledge.

McEnery, A. M., and Xiao, R. Z. (2007). Parallel and comparable corpora: What are they up to? In G. James, & G. Anderman (Eds.), *Incorporating Corpora: Translation and the Linguist* (Translating Europe). Clevedon, UK: Multilingual Matters.

Moreno, A.I. (2008). The importance of comparable corpora in cross-cultural studies. In Ulla Connor, Ed Nagelhout, and William Rozycki (eds), Contrastive Rhetoric: Reaching to Intercultural Rhetoric, pp. 25-41. Amsterdam: John Benjamins.

Parington, A. (2004). Corpus Linguistics: What It Is and What It Can Do. CULTUS. 32-54.

Richards, J., and Schmidt, R. (2002). Longman Dictionary of Language Teaching and Applied Linguistics. London: Pearson Education Linited.

Sultan, A. H. J. (2011). A contrastive study of metadiscourse in English and Arabic.ACTA Linguistica 5(1): 28-42.