Using Online Corpora Databases to Improve ESL Writing

Basem Y. Alkazemi[1*], Grami M. Grami[2]
[1] Umm Al Qura University, Makkah, Saudi Arabia.
[2] King Abdulaziz University, 80202, Jeddah, 21589, Saudi Arabia.

* Corresponding author. Tel.: 00966594238814; email: bykazemi@uqu.edu.sa
Manuscript submitted April 10, 2015; accepted November 20, 2015.
doi: 10.17706/jcp.11.5.374-379

Abstract: This paper summerizes the technical specifications and design principles of an online search tool we developed to help English as a Second Language (ESL) learners in certain aspects of writing. The problem we identified is the confusion caused by either first language interference (L1 interference) whereby literal translation erroneously combines with English words, or because certain words in English come together without a clear rule governing their combination such as collocations. The tool designed employs databases of authentic English texts (corpora) in addition to texts from trusted websites such as governmental and educational websites.

Key words: ESL writing, online databases, corpora.

1. Introduction

The concept of an online search engine that searches results within the massive corpora databases is not a novel idea. In fact, the underlying principle of most concordances currently available online is that. However, there have been certain shortcomings of these engines mainly because their search results cannot be filtered down according to location, origin, subject of study. Furthermore, many reliable results from reputable publications including – but not limited to – newspapers, books, legislative and educational print to name but a few.

Anyone who attempted to write in a foreign language will identify with the difficulties associated with such a demanding task. In fact, any assistance as far as composition is concerned should be of great value. What concordances provide is a reliable source for different accepted combinations of words. Many of these combinations are accepted not based on their syntactic of semantic virtues, but due to their collocations with each other. The concept of collocations is not readily understandable and for a non-native speaker can be especially difficult to comprehend.

In this technical paper, we discuss the bases of our project, the technical details of the search engine developed, and we identify the target users and how they can benefit from its results.

2. Literature Review

2.1. Writing Assistance Software

Ref. [1] and Ref. [2] opt for two different approaches in designing intelligent writing assistance programmes. The former, for instance, focuses primarily on what he calls 'text critiquing' which mainly addresses issues of grammatical accuracy. The latter, however, combines machine translation (MT) and monolingual writing. Again, the focus of this approach tends to be grammatical accuracy and style, and

although it provides a more comprehensive form of assistance to writers, issues such as word selection are still not fully investigated. [3] employed a statistical translation method (SMT), a well-known approach in electronic translation, to identify what they call mass errors in Chinese error corpus. The purpose was to identify common errors for later training purposes. Recent versions of word processing software such as Office 2010 provide limited feedback in terms of word selection. However, they do not provide examples in which a phrase is used or even results from different outlets.

It can be argued that a comprehensive word-combination checker is useful since very limited rules exist to help students decide which of many similar words is the most appropriate in their respective context. Furthermore, students can learn from their mistakes when errors tend to occur more often in terms of their linguistic background as [3] point out.

In other words, a more comprehensive word-checker should cover not only the accuracy of a combination of words but also that it follows the conventions of collocation and other types of phrases. The feedback should also be up to date and highly customizable, by which we mean students can check he level of formality and the origin of the text in question. Both requirements cannot be achieved using preinstalled word assistance programmes especially when they are not connected to the internet to allow for most recent results.

2.2. ESL Writing

There are many types ('genres' to use the technical term) of writing depending on the field of study or the context of discourse. For instance, some examples of genre in literature include drama, satire, tragedy, comedy, fantasy, folklore, horror, humour, poetry, mystery and science fiction. The classification goes further into subgenres such as black comedy and parody under the comedy heading. Similarly, academic genres are mostly discipline-specific, even if intertextual. The following are examples of academic genres: medicine, law, psychology, technology, philosophy, rhetoric and many others. Different writing genres can result in different translations of the same phrase; hence the need to identify which text belongs to which genre for successful translation to take place.

Another issue is collocation. As this is a problematic area for many L2 learners, even in the case of more advanced ones, especially in the lexical level of ESL writing. This view is supported by many experts in the field of education.

Linguists mention that collocation is part of lexical cohesion. It is also associated with corpus linguistics and can be defined as the syntactic/grammatical association of lexical items, quantifiable, textually, as the probability that there will occur at n removes (a distance of n lexical items) from an item x, the items a, b, c. In basic terms, collocation refers to certain words commonly used together which co-occur more often than only by chance. Examples of collocation in English include the use of verbs like 'do' and 'make' and the adjectives 'quick' and 'fast' with certain nouns, for instance one can say 'do your homework' and 'make a sandwich' but it is unusual to swap the verbs in these commands even if syntactically correct. The same applies to 'fast train' and 'quick shower' [4].

This is one reason why collocation is confusing, because 'do' and 'make' are almost synonymous to many ESL students who would assume, probably when applying L1 analogy, that they are interchangeable. Verb + noun collocation is possibly the most common type, but there are also verb + adverb (vividly remember), adverb + adjective (fully aware), adjective + noun (excruciating pain), and noun + noun (ceasefire agreement) collocations.

Finally, [5] also acknowledge English compound words, phrases, idioms and proverbs as being problematic areas in translation. Not just because a word-for-word translation would be incorrect, but also because these example tend to have various meanings, depending on the context. This makes them especially challenging

to the L2 writer.

2.3. Collocations in English

Linguists mention that collocation is part of lexical cohesion. It is also associated with corpus linguistics and can be defined as the syntactic/grammatical association of lexical items, quantifiable, textually, as the probability that there will occur at n removes (a distance of n lexical items) from an item x, the items a, b, c. In basic terms, collocation refers to certain words commonly used together which co-occur more often than only by chance. Examples of collocation in English include the use of verbs like 'do' and 'make' and the adjectives 'quick' and 'fast' with certain nouns, for instance one can say 'do your homework' and 'make a sandwich' but it is unusual to swap the verbs in these commands even if syntactically correct. The same applies to 'fast train' and 'quick shower' [4].

This is one reason why collocation is confusing, because 'do' and 'make' are almost synonymous to many ESL students who would assume, probably when applying L1 analogy, that they are interchangeable. Verb + noun collocation is possibly the most common type, but there are also verb + adverb (vividly remember), adverb + adjective (fully aware), adjective + noun (excruciating pain), and noun + noun (ceasefire agreement) collocations.

Teaching collocation to non-native speakers is important, unlike their native counterparts who acquired these examples spontaneously since they were children and throughout their daily lives. [6] mention some of the problems English learners may encounter due to L1 interference and vocabulary confusion in general and collocation in particular. [7] therefore believe that unless collocation is taught an important aspect of language will be overlooked.

How to teach collocation is a topic of paramount importance. Providing instructions in isolation is no longer an effective way of learning. Integrating it into reading and writing courses can be a much better alternative. Interestingly, researchers also indicate that the dichotomy between grammar and vocabulary is invalid and they argue that it is likely to be a misconception that learners have to study grammar thoroughly, then learn lots of words, after which they can say whatever they wish. This might indicate that teaching collocation in context rather than in isolation can achieve better results.

3. Concise Collocation Checker

Having identified the shortcomings of many writing assistance programs based on machine translation technologies and investigated areas where ESL writers are likely to make errors in; we in effect have laid the foundations on which a better writing assistance program can be built. Work has started with one simple technique of finding exact results on available online text to confirm whether a combination of 2, 3 and 4 words exists or not. More parameters were suggested later on to judge the accuracy and formality of the phrase depending on the text is appears on, its origin and locality.

The theoretical framework has been implemented into a feasible prototype which has the capability of searching for specific word combinations and providing up-to-date results from the available databases of search engines. Many desired features have not yet been implemented. However, they are being developed and will be included in the upcoming versions of the tool. The main difficulty - as ever - was the limited time available and the lack of resources. Despite this, the teams are determined to improve the original design, even if at a slower rate than originally planned.

The programmers responsible for developing consisted of four final-year computer science students at UQU University working on the project as part of their graduation degree. The features of the tool were divided into two categories - functional and non-functional. This division helped the team allocate tasks to individuals and maintain a timely work plan. It also worked as a basis for our testing plan which is laid out in

this report.

3.1. The Technical Requirements

The functional requirements identified by the team consisted of eight divisions which are further divided into four releases, all designed to reflect the principles of design and functionality expected of the software. They are described in the accompanied graphs.

3.2. The Non-Functional Requirements

The non-functional requirements were identified as follows:

Layout

The early releases of the tool will have web-based application orientation. However, the average user should expect it to be integrated in a word-processing program such as Open Office. As a result, look-and-feel requirements are not significant at the moment, as the tool works in the background of the word-processor. The only consideration of this non-functional requirement is in the pop-ups related to user recommendations. The look-n-feel of these pop-ups will be inherited from that of the word-processor.

Usability

The tool will be integrated as a plug-in to a word-processor and works as a background process which should help users make checks seamlessly and with very little difficulty.

Performance

The availability of a predetermined database (DB) will eventually enhance the overall performance and response time of the system in the event of Internet connectivity disruption.

Portability

At a later stage of this project we are going to define the API of the tool in order to facilitate portability and integration. We are going to design our tool to be modular enough to work as a stand-alone application, in addition to working as a plug-in to a word processor.

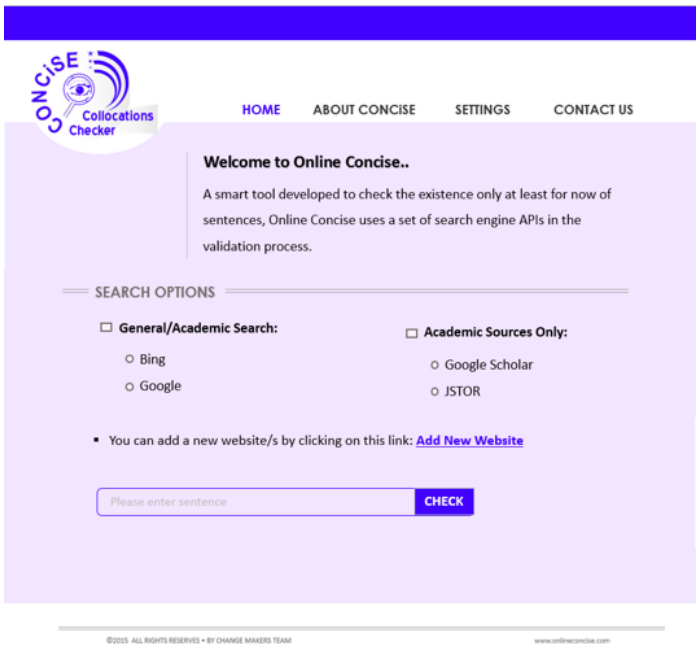

Fig. 1. The homepage screenshot.

User Interface

Ideally, the project is going to deliver a number of releases to potential customers. Our ultimate goal is to design a tool integrated as a plug-in into word processors as shown in Fig. 1. Each release is planned to

feature a slightly different user interface, courtesy of any added features.

Fig. 2. About concise.

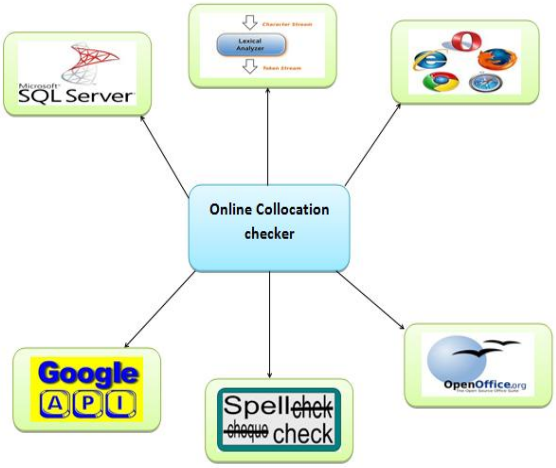

Fig. 3. Software interfaces.

The user interface follows a search engine layout that allows the user to type a sentence and search Google for matching results. The retrieved results will be displayed in a box underneath the search field. It is going to display statistics of the number of results obtained from a pre-defined list of sites. So, the user can be advised of the correctness of the sentence they have typed. Another feature is the ability to select different search engines.

This customization is more than usually possible with typical concordance search tools. They can also be adjusted to suit users specific field of interest e.g. if they are writing a political article they can choose texts from political sites.

Software Interfaces

The development of the tool is in accordance with the MVC architectural pattern which seems to be a suitable decision given the size and potential of the software.

Concise interacts with various systems including search engines, word processors (e.g. Open Office), SQL database, Lexical analyzer tool, and spelling and grammar checkers.

4. Conclusion

Although using search engines to check the accuracy and acceptability of certain word combinations is not in itself a new concept in ESL writing, the tool we developed does in fact combine the functionalities of corpora concordances in addition to search pools from academic and other trusted sources. In theory, word-combination based on collocations is problematic and in most cases the best solution for ESL students is to check whether a phrase they write does in fact exists and to which effect. The search engine in its early iteration is very promising indeed. We however acknowledge the need for further research and development as well as integration with various platforms to achieve its full potential. For instance, we intend to integrate the search engine directly with word processing software to allow users to have instant feedback along with examples and comments.

Acknowledgment

The Institute of Research and Revival of Islamic Heritage at Umm Al Qura University, Makkah, Saudi Arabia, has awarded a grant of SAR 115,200 toward developing this project (grant number 43308015).

References

[1] Heidorn, G. (2000). Intelligent Writing Assistance. *Handbook of Natural Language Processing*, 181-207.

[2] Brockett, C., & Dolan, W. (2010). *U.S. Patent No. 7,752,034*. Washington, DC: U.S. Patent and Trademark Office.

[3] Brockett, C., Dolan, W. B., & Gamon, M. (2006). Correcting ESL errors using phrasal SMT techniques. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 249-256). Association for Computational Linguistics.

[4] Guo, S., & Zhang, G. (2007) 'Building a customised Google-based collocation collector to enhance language learning'. *British Journal of Educational Technology*, *38(4),* 747–750.

[5] Lindstromberg, S. (2003). My good-bye to the lexical approach. *Humanising Language Teaching*, *5(2).*

[6] Swan, M., & Smith, B. (2001). *Learner English: A Teacher's Guide to Interference and Other Problems*. Cambridge: Cambridge University Press.

[7] Lewis, M., & Conzett, J. (2000). *Teaching Collocation: Further Developments in the Lexical Approach.* Hove: Language Teaching Publications.

Basem Y. Alkazemi was born in Makkah, Saudi Arabia. He received his PhD in software engineering from Newcastle University in the UK in 2009. He is an associate professor of computer science at Umm Al Qura University, Saudi Arabia. He received his PhD from Newcastle University in the UK and went on to work on many granted projects and publications.

Grami Mohammad A Grami was born in Jeddah, Saudi Arabia, 1980. He received his PhD in education and applied linguistics from Newcastle University in 2010. He is a teaching assistant at King Abdulaziz University in the Department of European Languages. Along with Dr. Alkzemi, he published many articles with focus on technology and education.