2 ICLEHI 2015 60 Roslina Abdul Aziz

Building the Malaysian Corpus of Financial English (MaCFE)

Roslina Abdul Aziz*, Noli Maishara Nordin, Mohd Rozaidi Ismail,
Norzie Diana Baharum, Roslan Sadjirin,
Universiti Teknologi MARA Cawangan Pahang,
26400, Bandar Tun Abdul Razak Jengka, Pahang, Malaysia
*Corresponding Author: leenaziz71@gmail.com

ABSTRACT

The paper describes the process involved in developing the Malaysian Corpus of Financial English (MaCFE); a specialized online corpus, which will contain a wide range of written texts obtained from various financial institutions in Malaysia. This corpus will also be the first sub-corpus to contribute to the development of the Malaysian Corpus of English for Specific Purposes. It is built using the corpus linguistics methodology involving four main research activities; (i) Digital Database Construction, (ii) Annotation, (iii) Web Interface Development and (iv) Data Web Release. Once completed the corpus will not only be an important source of reference for research, teaching, and learning of ESP in Malaysia, but it will also serve as a major resource for the training of future financial services professionals.

Keywords: Corpus Linguistics, Specialized corpus, Financial English, English for Specific Purposes (ESP), Malaysian Corpus of Financial English (MaCFE)

Introduction

Specialized corpora such as the English for Specific Purposes (ESP) corpora are not easily accessible. To date, only a small number of specialized corpora can be publically accessed online. They include the Michigan Corpus of Academic Spoken English (MICASE), the Corpus of London Teenager Language (COLT), the Hong Kong Engineering Corpus (HKEC) and the Hong Kong Financial Services Corpus (HKFSC). The majority of other specialized corpora (for instance the Cambridge Business English Corpus, the Cambridge Legal English Corpus, and the Cambridge Academic English Corpus) are only accessible through purchase or subscription (Warren, 2010). The growth of corpus-informed pedagogy (CiP) as a complement to the more traditional approach to language teaching, the burgeoning use of corpora in language teaching for general English and for ESP courses, and the important contribution of corpora as the resource and reference for the design and development of specific language curriculum (Skorczynska, 2010; Phoocharoensil, 2012; Warren, 2010; Yoon, 2008), call for more specialized corpora to be developed and published. The scarcity of publically available ESP corpora worldwide could impede the progress in the research and development in ESP. In addition, the specialized corpora available online only cover a small area of ESP (Warren, 2010) and they consist of language data of very specified discourse communities, for instance HKEC and HKFSC which consist of financial and engineering language used in Hong Kong. These corpora could not possibly be fully incorporated in the Malaysian ESP contexts or be utilized as a resource and reference for the purpose ESP teaching and learning in this country. To date there has been no record of specialized corpora being developed in Malaysia. There is, therefore, the need for an ESP corpus to be developed that would comprise the English used by specific discourse communities in Malaysia. Based on this need, this study aims to build a corpus that can cater

for ESP research and teaching in Malaysia. This project begins with the development of its first sub-corpus, which focuses on the English used by financial services in Malaysia. The sub-corpus will be known as the Malaysian Corpus of Financial English or MaCFE.

Purpose of the study

The purpose of this study is to develop a specialized online corpus, which contains a wide range of written texts obtained from the financial institutions in Malaysia.

Research questions

The study intends to answer the following research questions:

1. What text types represent the English used by the financial services in Malaysia?
2. What processes are involved in the design and development of MaCFE?
3. What processes are involved in publishing MaCFE online?

Literature Review

Corpus-informed Pedagogy (CiP)

The integration of corpus in the context of L2 learning and teaching has become increasingly appealing in recent years (Gaskel & Cobb, 2004; Yoon & Hirvela, 2004; Yoon, 2008, 2011; Phoocharoensil, 2012). The ability to simultaneously integrate language skills such as vocabulary, grammar, writing and reading, makes corpus consultation an attractive complement to the traditional method of language teaching and learning. The corpora, which are the collection of massive language database from multiple resources, offer learners a rich exposure to the genuine language (Thurstun & Candlin, 1998, Yoon & Hirvela, 2004), the kind that they will most likely encounter and eventually use outside the classroom. The exposure to genuine language use within its contexts can enrich learners' understanding and repertoire in the use of the target language (Yoon & Hirvela, 2004), thus contributing to their overall growth in it.

The integration of corpora with the language pedagogy can also promote inductive learning, where learners take central stage in their own language learning process (Johns, 1991; Todd, 2001; Lee & Swales, 2006, O'Sullivan, 2007). In his pioneering work in direct corpus integration with the teaching of vocabulary and grammar, a language learning approach he termed as data-driven learning (DDL), Johns (1991) introduced the learners as language researchers. The learners perform the role of research workers (Johns, 1991) or language explorers/travelers (Benardini, 2004) who are actively engaged in analyzing and discovering language patterns or rules from the authentic data available from the corpora. The approach requires the learners to be directly confronted with the language, analysing and deriving the language patterns instead of just being passive receivers at the end of the teaching and learning continuum; receiving rules and patterns from the teachers or grammar books and language references. This exploratory nature of DDL can be highly "motivating and highly experiential" (Ketterman, 1995:10) for the learners and can be much more interesting and rewarding than being taught about language (Phoocharoensil, 2012). For L2 learners to achieve high level of proficiency in the target language, they need to be actively involved in the acquisition process; researching the language, developing and testing hypotheses, and applying language patterns in genuine communication contexts. All these language-learning expectations can be fulfilled through the inclusion of corpora.

Another reason that makes CiP more interesting and attractive is the computer-based nature of corpus consultation. The use of computer provides L2 learners greater exposure to the target language and allows for greater opportunities for interaction with it (Yoon and

Hirvela, 2004). Furthermore, the abundance of computer-based artifacts such as the internet and hypertext, and the availability of online corpora (Bank of English sampler, Collins COBUILD sampler and Michigan Corpus of Academic Spoken English) allow for limitless access to the target language texts (Sun, 2000; Conrad, 2001). The use of computers also allows L2 learners more freedom; they not only gain access to the endless texts to choose from but are also given the upper hand to select texts that would have "the greatest linguistic value" relative to the needs and requirements of their studies (Yoon & Hirvela, 2004: 261). By comparison, the opportunities to interact and work with texts specified to the learners' discoursal domains can be more linguistically rewarding than working with the general texts available in the textbooks. Most importantly, the direct access to the vast arrays of language resources immediately and readily available at any given time, encourages the learners to use these resources outside the classrooms, thus, providing ample opportunities for the learners to be autonomous in their own learning and become more independent language learners (Yoon & Hirvela, 2004).

For the reasons discussed above corpus-informed language pedagogy has gained support and momentum in recent years. A number of empirical studies on the effects of corpus applications in the teaching of lexico-grammatical patterns, grammar, vocabulary, collocations (Todd, 2001; Chambers & O'Sullivan, 2004; Kennedy & Miceli, 2001; Phoocharoensil, 2012) and writing (Yoon & Hirvela, 2004, Yoon, 2008, Kennedy & Miceli, 2001, 2002; Gaskell & Cobb, 2004; Chambers & O'Sullivan, 2004) have been conducted. The findings of these studies revealed that the inclusion of corpus has proven effective in improving learners' knowledge in vocabulary (Todd, 2001; Miceli & Kennedy, 2001), grammar (Todd, 2001; Phoocharoensil, 2012) and general writing skills (Kennedy & Miceli, 2001, 2002; Yoon & Hirvela, 2004; Gaskell & Cobb, 2004; Chambers & O'Sullivan, 2004, Yoon 2008).

Corpora in English for Specific Purposes (ESP)

Corpora can be significant language sources for teaching and learning of English for Specific Purposes (ESP). One example is the use of a specialized corpus by the Iowa State University to help their students write research papers in their designated fields (Cortes, 2007). Another example is the use of Contemporary Written Italian Corpus (CWIC), a specialized corpus comprising of private and official letters, private and official e-mails, letters to the editors, magazine articles and film reviews written by native speakers of Italian. The corpus was utilized in developing less advanced learners' writing skills in Italian by means of corpus-based error correction and content and vocabulary enriching activities (Kennedy & Miceli, 2001).

In addition, corpora are also significant reference tools to access authentic language used in a target discourse community. Many corpus-based studies conducted in the area of ESP (Candlin, Bhatia, & Jensen, 2002; Hyland, 1994; Paltridge, 2002; Skorczynska, 2010; Swales, 2002; Williams, 1988) have come to the same conclusion, that published materials for the teaching of ESP and EAP are not able to reflect the real academic or business discourse. Williams (1988) criticized the unnatural sounding linguistic exponents in published business English textbooks, which according to the researcher fail to showcase a high degree of linguistic complexity in real business meetings. Skorczynska (2010) reported significant differences in both the type and use of metaphors listed in published business English textbooks (MacKenzie, 1997) in comparison with those found in a specialized corpus of written business English. Corpora can provide the much needed sources of authentic language used in a target discourse community and more importantly these sources can be manipulated by the teachers as the resources and tools for the teaching of ESP or EAP.

Moreover, investigations on ESP corpora are highly valued for their ability to provide course designers and ESP practitioners of salient lexico-grammatical features, typical choice of words (frequency), meaning nuances of near-synonyms and appropriate use of collocations of a target discourse community and the genre structures important in designing specific language curriculum (Flowerdew, 1993). An excellent example of this is the use of the Cambridge English Corpus in the publication of ESP reference books such as the Cambridge English for Engineering, the Cambridge English for Human Resources, and the Cambridge English for Nursing, to name a few from a series of ESP reference books produced by Cambridge University (http://www.cambridge.org).

The literature reviewed highlights the significant contribution of specialized corpora in the development and advancement of ESP worldwide (Skorczynska, 2010; Phoocharoensil, 2012; Warren, 2010; Yoon, 2008). Nevertheless, presently there are only a handful of specialized corpora that can be accessed online and be fully utilized for the purpose ESP research, teaching and learning. More importantly, to this date none of these specialized corpora contain language data representing the English used by the various discourse communities in Malaysia. These reasons, thus, have ignited the motivation to develop and publish an ESP corpus that could be utilized by researchers and educators in the planning, developing and teaching of ESP in Malaysia.

Methodology

MaCFE is built following the methodology of corpus linguistics at its current state. Adopting the Aksan and Aksan (2009) work packages, the building of MaCFE involves four major development packages i.e. (1) Electronic Database Construction, (2) Annotation, (3) Interface Development and (4) Pre-release Control. (Please refer to Appendix A to view the workflow proposed for the building of MacFE.)

Electronic Database Construction

External criteria. The first major process involved in the construction of the electronic database is data compiling, which would commence once the external criteria and text types for the corpus are determined. A corpus according to Sinclair (2005) comprises of a collection of carefully selected pieces of language text in electronic form, which are compiled according to the external criteria that represent, as far as possible, a language or language variety as a source of data for linguistic research. The external criteria that are most common, according to Sinclair (2005) include:

i. the mode of the text; whether the language originates in speech or writing, or perhaps nowadays in electronic mode;
ii. the type of text; for example if written, whether a book, a journal, a notice or a letter;
iii. the domain of the text; for example whether academic or popular;
iv. the language or languages or language varieties of the corpus;
v. the location of the texts; for example (the English of) UK or Australia;
vi. the date of the texts.

Sinclair (2005:1-16)

MaCFE is proposed to be built according to the external criteria specified above (Sinclair, 2005) and the details of the criteria are presented in Table 1.

Table 1
External Criteria for MaCFE

Mode	Written	
Text Type	Printed/Electronic	Annual Reports, Brochures, Codes of Practice, Corporate Announcement, Circulars, Product Descriptions, Product Reports, Interim Reports, Media Releases, Ordinance, Prospectuses etc.
Domain	Financial	Banking: Commercial Banks, Islamic Banks, International Islamic Banks and Investments Insurance: Conventional Insurance Life and General Business, Life Business only, Life and General Reinsurance Business, Life Reinsurance Business Takaful Takaful Operators, Retakaful Operators and International Takaful Operator (source: Bank Negara Malaysia at http://www.bnm.gov.my, 2014)
Language	English	Authors: Non-Native Speakers of English
Location	Malaysia	
Date	2010-2014	

Text types. As for the text types, MaCFE would mainly consist of written texts that could either be in the printed or electronic form. The data would be collected mainly from the financial services institutions in Malaysia. Prior to data collection, the researchers are required to (i) consult advice from experts in the field of financial services to determine the text types representing the field and (ii) set a comprehensive text typology for the study. For this study, the data would consist of the text types adapted from Warren (2010) as summarized in Table 2 below. It is important to highlight that the text types established by Warren (2010) will only serve as a guideline as the actual text types will be determined following the advice obtained from the financial services experts.

Table 2
Proposed Text Types for MaCFE

Text Type	Text Type
Annual Reports	Insurance Product Descriptions
Brochures	Investment Product Descriptions
Bank Service Charges	Model Agreements
Codes of Practice	Media Releases
Corporate Announcement	Ordinances
Circulars	Procedures

Fund Descriptions	Principles
Fund Reports	Prospectus
Factsheets	Rules
Guidelines	Results Announcements
General Meeting	Standards
Insurance Policies	Speeches
Interim Reports	

Warren (2010)

Once the data collection process is completed, the data would then have to be digitalized to transform them into electronic/machine-readable texts. The digitalizing process involves scanning and keyboarding printed and written data and saving each as a text file to be categorized and stored.

The computerized data need to go through a cleaning process, which include (i) removing unnecessary elements from the original documents for instance author names, graphics, tables, figures and other texts or information irrelevant to the documents, (ii) in addition, for texts obtained from the internet it is also necessary to remove the HTML tags, comments and scripts and menu from the original web.

Annotation

MaCFE would require two major types of annotation; meta-linguistic and part of speech (POS) annotation. The meta-linguistic information such as the title, mode, publication date, number of words will be added to all texts. The British National Corpus (BNC) model will be used for this purpose. POS annotation on the other hand involves tagging the lexemes in the data with a commercial automatic POS tagger available in the market.

Web Interface Development

The corpus would be published online using the IMS Open Corpus Workbench (CWB). CWB is a collection of open-source tools for managing and querying large text corpora (ranging from 10 million to 2 billion words) with linguistics annotations. Corpus Workbench (CWB) initially designed at the University of Stuttgart, is a widely used software architecture for corpus analysis. It contains a set of powerful tools for indexing, managing and querying very large corpora with multiple layers of word-level annotation. Its central component the Corpus Query Processor (CQP) is an extremely efficient concordance system. Both CWB and CQP are very commonly used as the back-end for many web corpus interfaces for instance the British National Corpus (Evert & Hardie, 2013).

The query criteria will be based on the existing features provided by BNCweb-CQP edition. The interface will include features to produce concordance output, specification of the categories of texts, section of texts, facility to access meta-textual information and frequency data. The visual designs will be user-friendly following the existing Web corpus such as BNCweb. Cross-platform compatibility will focus on three area: (i) Character set, texts in the corpus will be saved in UTF-8 character code, (ii) Platform-free corpus interface will be constructed in terms of w3c standards and can be accessed through any web-browser on any platform and (iii) Result set, results can be used on any platform and they will be compatible with other software.

Data Web Release

The pre-release of MaCFE involves two beta and two candidate releases for local, national and international testing. Version 1 will be released for local testing, which involves bringing out the corpus on the local network. The same version is then released for national and international testing. Upgrades will be applied for Version 1 after receiving feedbacks after each trial.

Conclusion

The completion of the MaCFE would mark an important chapter in the corpus linguistics field in Malaysia. It will be the first specialized corpus to be built in Malaysia and will be among the very few in the world that could be freely accessible online. It will also pave an important path for the creation of other specialized corpora that would contribute to the building of a larger and more comprehensive Malaysian Corpus of English for Specific Purposes or MaCESP.

References

Aksan, Y., & Aksan, M. (2009). Building a national corpus of Turkish: Design and implementation. *Working Papers in Corpus-based Linguistics and Language Education,* No. 3, 299-31. Tokyo: TUFS.

Aston, G. (1997). Small and large corpora in language learning. In B. Lewandowska-Tomaszczyk & J. P.Melia (Eds.), *Practical applications in language corpora* (pp. 51-62). Lodz, Poland: Lodz University Press.

Candlin, C., Bhatia, V., & Jensen, C. (2002). Developing legal writing materials for English second language learners: Problems and perspectives. *English for Specific Purposes,* 21, 299–320.

Bernardini, S. (2004). Corpora in the classroom: an overview and some reflections on future developments. In J. M. Sinclair (Ed.), *How to use corpora in language teaching,* Amsterdam, Netherlands: John Benjamins. 15–36.

Chambers, A. and O'Sullivan, Í. (2004). Corpus consultation and advanced learners' writing skills in French. *ReCALL,* 16(1): 158–172.

Conrad, S. M. (2001). Variation among disciplinary texts: A comparison of textbooks and journal articles in biology and history. In S. M. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies*. Harlow: Longman.

Cortes, V. (2007). Teaching lexical bundles in the discipline: An example from a writing intensive history class. *Linguistics and Education,* 17 (4). 391-406.

Dodd, B. (1997). Exploring a corpus of written German for advanced language learners. In: Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds.) *Teaching and Language Corpora*. London, New York: Longman, 131–145.

Evert, S. & Hardie, A. (2012). Twenty-first Century Corpus Workbench: Updating a Query Architecture for the New Millenium. *Proceedings of Corpus Linguistics* 2011. Birmingham, UK.

Flowerdew, J. (1993). An educational, or process, approach to the teaching of professional genres. *English Language Teaching Journal,* 47 (4). 305-316.

Fox, G. (1998). Using corpus data in the classroom. In: Tomlinson, B. (ed.), *Materials Development in Language Teaching*. 25–43, Cambridge: Cambridge University Press.

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System,* 32, 301–319.

Hyland, K. (1994). Hedging in academic writing and EAP course books. *English for Specific Purposes,* 13, 239–256.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *ELR Journal,* 4, 1–16.

Johns, T. (1994). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.), *Perspectives on pedagogical grammar*. 293-313, Cambridge: Cambridge University Press.

Johns, T. (1997). Contexts: the background, development and trialling of a concordance-based CALL program. In: Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds.), *op. cit*., 100–115.

Kennedy, G. (1987). Expressing temporal frequency in academic English. *TESOL Quarterly,* 21, 69–86.

Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology,* 5(3), 77–90.

Kettemann, B. (1995). On the use of concordancing in ELT. *TELL & CALL ,* 4, 4-15.

Kettemann, B. (1996). Concordancing in English language teaching. In: Botley, S., Glass, J., McEnery,T. and Wilson, A. (eds.), *Proceedings of Teaching and Language Corpora,* 4–16, 1996. Lancaster: University Centre for Computer Corpus Research on Language.

Mackenzie, I. (1997). *Management and marketing with mini-dictionary of 1000 common terms*. Hove: LTP.

O'Sullivan, Í. and Chambers, A. (2006) Learners' writing skills in French: corpus consultation and learner evaluation. *Journal of Second Language Writing,* 15(1): 49–68.

O'Sullivan, I. (2007). Enhancing a process oriented approach to literacy and language learning: The role of corpus consultation literacy. *ReCALL* 19 (3), 269-286.

Paltridge, B. (2002). Thesis and dissertation writing: An examination of published advice and actual practice. *English for Specific Purposes,* 21, 125–143.

Phoocharoensil, S. (2012). Language corpora for EFL teachers: An exploration of English grammar through concordance lines. *Procedia – Social and Behavioral Sciences* 64, 507 – 514.

Sinclair, J. (2005). Corpus and text-Basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice*. Oxbow Books: Oxford. 1–16. Retrieved February 20, 2015, from http://ahds.ac.uk/linguistic-corpora/.

Skorczynska, H. (2010). A corpus-based evaluation of metaphors in a business English textbook. *English for Specific Purposes,* 29, 30–42.

Stevens, V. (1995) Concordancing with language learners: Why? When? What? *CÆsLL Journal,* 6(2): 2–10. http://www.eisu.bham.ac.uk/johnstf/stevens.htm.

Sun, Y. -C. (2000). *Using online corpus to facilitate language learning*. Paper presented at the Annual Meetingof the Teachers of English to Speakers of Other Languages, British Columbia, Canada.

Swales, J. (2002). Integrated and fragmented worlds: EAP materials and corpus linguistics. In J. Flowerdew (Ed.), *Academic discourse*, 150–164, London: Longmans.

Thurstun, J., & Candlin, C. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes*, 17(3), 267–280.

Todd, W. R. (2001). Induction from self-selected concordances and self-correction. *System* 29, 91-102.

Yoon, H. & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, 13, 257-283.

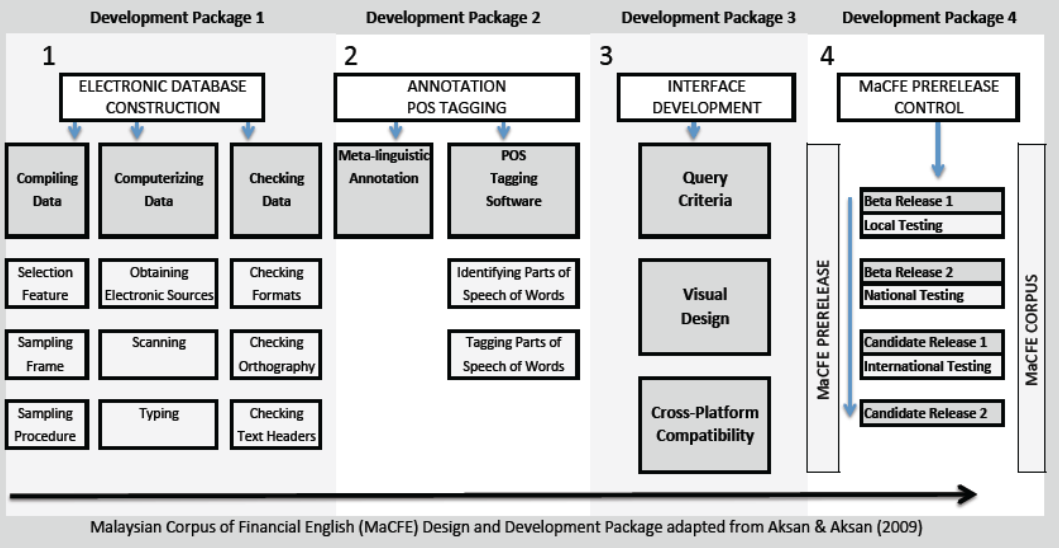
Yoon, H. (2008). More than a linguistic reference: the influence of corpus technology on L2 academic writing. *Language Learning & Technology,* 12(2), 31–48.

Yoon, H. (2011). Concordancing in L2 writing class. An overview of research and issues. *Journal of English for Academic Purposes,* 10, 130-139.

Williams, M. (1988). Language taught for meetings and language used in meetings: Is there anything in common? *Applied Linguistics,* 9 (1), 45-58.

Warren, M. (2010). Online corpora for specific purposes. *ICAME Journal,* 34, 169-188.

APPENDIX A



Malaysian Corpus of Financial English (MaCFE) Design and Development Package adapted from Aksan & Aksan (2009)