Development of a Mesolectal Malaysian English Corpus

Christina Ong Sook Beng
Universiti Tunku Abdul Rahman, Kampar Campus

ongsb@utar.edu.my

Introduction

To date, there are more learner corpora in Malaysia compared to general and specialised corpora. They have undoubtedly assisted in the discovery of various grammatical errors made by Malaysian students. Another relatively popular type of corpus is newspaper/print media, specialised corpus that has been used to analyse grammatical variation. These imply the negligence of general corpora representing Malaysian English with the exception of two ongoing corpora—International Corpus of English-Malaysia (ICE-M'sia) and Malaysian Academic Spoken English Corpus (Siti Aeisha & Hajar, 2014). Davies and Fuchs (2015) stressed that these corpora are created aiming to investigate certain trends by individual researchers which are not able to cater to researchers of varieties of English, in this case Malaysian English.

A paradigm shift from general corpus to big-data-based corpus for linguistic studies was observed since the launch of Corpus of Contemporary American English—COCA (approximately half a billion words) in 1990 (Davies, 2009), followed by GloWbE and the most recent iWeb corpus consisting of 1.9 billion and 14 billion words respectively as stated in English-Corpora.org. Using web data to facilitate the study of language variation has yielded many interesting findings owing to easy accessibility of countless webpages encompassing contemporary formal and informal English texts from any country. Consequently, general corpus like ICE with one-million-word is side-lined as proven by Davies and Fuchs (2015) who asserted that ICE is inadequate for in-depth studies on morphological, lexical, syntactic and semantic variations. Hundt, Nesselhauf, and Biewer (2007) substantiated it by pointing out scarcity of data in standard corpora when linguistic items investigated are rare or too new. To keep up with the evolution of corpus linguistics and the need for big corpora to investigate language variation, Corpus of Malaysian English Forum, Malaysia first corpus using web sources will be built.

This study reviews challenges and chances involved in developing Corpus of Malaysian English Forum (CMEF), a general corpus representing mesolectal Malaysian English to facilitate studies on nativisation of New Englishes. Prior to that, its composition and parameters governing the creation of CMEF are discussed.

The Composition of CMEF

Following the prominence of informal language in general corpora, changes are observable when spoken data or informal language is more prominent in many renowned corpora used to investigate language in New Englishes. One of the possible and suitable sources of informal language is computer-mediated communication (CMC) texts, specifically the Internet forum. Analysing computer-mediated writing is essential in the studies of varieties of English from the sociolinguistics view point (Mair, 2011).

Known as one of the liveliest forums in Malaysia containing discussions about various topics (Goh, 2014), Lowyat.Net is deemed suitable for gathering data to create CMEF. Lowyat.Net consists of 9 main sections. To have a decent selection covering various topics, all

9 sections alongside several sub-forums with various threads are included in the corpus as can be seen in Table 1.0. Following Hundt, et al.'s (2007) claim about having 100 million words as the standard size of modern corpora, CMEF follows suit. The average length of a sub-forum is roughly 1 million words except for threads under LYN community project, and classifieds which recorded fewer amount of words while the real world issues sub-forum under roundtable discussion recorded 30% more words than the average length of other sub-forums. The differing text lengths in CMEF coincides with Meyer (2002) who supported the inclusion of different kinds of text in corpora instead of longer texts (as cited in Clancy, 2010).

Table 1.0 Sub-Forums from Lowyat.Net (LYN) included in Corpus of Malaysian English Forum

Main sections	**(%)**
Front Desk	0.67
Computers	6.43
Special Interest	6.33
Roundtable Discussions	52.58
Entertainment	5.37
Lifestyle	23.34
LYN Community Projects	0.33
Classifieds	0.93
Trade Zone	4.02
Total	**100, 111, 842 words**

CMEF covers a wide range of topics – from technology to social issues and trading activities. To accommodate researchers of New English, sub-forums which are used as platforms to share images (e.g. photography, digital imaging & video) or to showcase artwork (e.g. arts & design) and containing discussions which are too niche (e.g. games – call of duty) are excluded.

Data of this corpus retrieved from the Internet forum should possess four characteristics of CMC which are interactive, international, interested, and intertextual (Richardson, 2001), except for the second characteristic as Lowyat.Net consists of mainly Malaysians. This is proven when its founder, Vijandren Ramadass claimed Lowyat.Net as a website disseminating information about gadget prices initially and it flourished into a forum focusing on issues happening in Malaysia later on (Goh, 2014). CMEF also meets three parameters of basic corpus highlighted by Claridge (2007) namely interactive (dialogic and polylogic forms), interested (not thematically restricted) and intertextual (evidence of parts or all previous messages quoted repeatedly). Most forums including Lowyat.Net depicts conversational-laden characteristic which can be associated to informal language or the spoken form of Malaysian English. Mair (2011) affirmed that forums contain more vernacular features than face-to-face conversations and that is recognised to promote identity construction. Therefore, texts gathered from Lowyat.Net to create CMEF are expected to represent the mesolectal sub-variety which is also regarded as the best representation of Malaysian English by Richards (1979), Platt and Weber (1980), and Baskaran (2005), ultimately foregrounding the Malaysian identity.

Challenges and Chances of Developing CMEF

Deciding the size of CMEF is quite complex. It is generally believed that the bigger a corpus is the better. Nonetheless, it is crucial to acknowledge Kennedy's (1998) observation that no matter how big a corpus is, it will never be able to capture all the output produced by the users of a language in a day (as cited in Tan, 2013). Despite the size of CMEF which probably

comprises a mere 10% of Lowyat.Net, it is adequate when instances of items investigated exemplifying this sub-variety can be generated. Another challenge in the creation of CMEF is the absence of spoken data. Unlike other forms of CMC like emails and chatrooms which are confined to a few individuals or at most a few thousand subscribed participants, forums according to Claridge are dialogic or polylogic sometimes and they are completely public (2007, p. 87). Because of the conversational nature of forums, they can be associated with oral linguistics features. This leads to another challenge that is the risk of including English language by non-Malaysians in the forum. It is unavoidable because similar to the anonymity of blog authors issue in GloWbE (Davies & Fuchs, 2015), CMEF shares the same risk because there might be non-Malaysians who have posted enquiries or responded in Lowyat.Net.

Undoubtedly, the World Wide Web offers accessibility of countless webpages encompassing both formal and informal English texts which are relatively current from any country in the world. Most formal texts in CMEF can be detected in sub-forums like education essentials, jobs and careers, and property talk under the fourth main section, roundtable discussions while texts in the remaining main sections are relatively informal. Besides easy accessibility, having control over what goes into the corpus and enabling searches which are impossible to run on raw web data is an advantage (Hundt, et al., 2007). For CMEF, as stated earlier, threads from all the sub-forums are included and this according to Claridge (2007) can ensure the representation of speakers from diverse backgrounds (but confined within Malaysia). In addition to the above-mentioned reasons, below are the chances for the creation of Malaysian general corpus using web sources:

i) the size of corpus created using web data will be relatively bigger compared to existing Malaysian corpora in English so that it can offer more examples of constructions which are non-frequent in specialised and general corpora.

ii) the texts gathered online will definitely be more updated and would reflect contemporary culture (Fletcher, 2011). As Lowyat.Net was founded in the early 21st century, the language used in the forum definitely reflects the most current linguistic scenario in Malaysia.

Conclusion

The need for developing CMEF, a general corpus representing Malaysian English and the reasons for extracting texts from Lowyat.Net, a forum which carries Malaysian identity have been described. To reiterate, three challenges encountered during the development of CMEF are: i) deciding its size; ii) countering absence of spoken data and; iii) ensuring its users are Malaysians. On the contrary, the chances are: i) its size is definitely bigger than other English corpora in Malaysia owing to the easy accessibility of the world wide web; ii) a decent selection of topics is included to ensure a balanced representation of mesolectal sub-variety and; iii) the texts retrieved from Lowyat.Net are certainly up-to-date. Supported by Loureiro-Porto, who claimed careful compilation of big data corpora is too attractive a source of material to ignore (2017, p. 468), it has nourished the study of languages in recent years. Although CMEF is not as huge as GloWbE or iWeb, it is believed to be able to yield interesting findings for nativised linguistic items in Malaysian English. Parallel with Claridge's (2007) belief in forum providing more updated linguistic variation compared to language represented in other corpora, the creation of CMEF is timely to allow research on nativisation of Malaysian English particularly grammatical variation to be conducted.

References

Baskaran, L. M. (2005). *A Malaysian English primer: Aspects of Malaysian English features.* Kuala Lumpur: University of Malaya Press.

Clancy, B. (2010). Building a corpus to represent a variety of a language. In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (80-92). United Kingdom: Taylor & Francis.

Claridge, C. (2007). Constructing a corpus from the web: Message boards. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web* (87-108). Amsterdam: Rodopi.

Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9-billion-word Global Web-based English Corpus (GloWbE). *English World-Wide, 36*(1), 1–28.

Davies, M. (2009). The 385+ million word corpus of contemporary American English (1990- 2008+).: Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*(2), 159 – 190.

Fletcher, W.H. (2011). Corpus analysis of the World Wide Web. In C.A. Chapelle (Ed.) *Encyclopedia of applied linguistics* (1339–47). Oxford: Wiley-Blackwell.

Goh, G. (2014). *Digerati50: Building online communities, the 'content way'*. Retrieved from https://www.digitalnewsasia.com/digital-economy/digerati50-building-online-communities-the-content-way

Hundt, M., Nesselhauf, N., & Biewer, C. (2007). *Corpus linguistics and the web*. Amsterdam: Rodopi.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London and New York: Longman.

Loureiro-Porto, L. (2017). ICE vs GloWbE: Big data and corpus compilation. *World Englishes, 36*(3), 448-470.

Mair, C. (2011). Corpora and the new Englishes: Using the 'Corpus of Cyber-Jamaican' (CCJ) to explore research perspectives for the future. In F. Meunier, S. de Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 209–236). Amsterdam: John Benjamins.

Meyer, C. (2002). *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

Platt, J., & Weber, H. (1980). *English in Singapore and Malaysia: Status, features, functions*. Kuala Lumpur: Oxford University Press.

Richards, J. C. (1979). Rhetorical and communicative styles in the new varieties of English. *Language Learning, 29*(1), 1-25.

Richardson, K. (2001). Risk news in the world of Internet newsgroup. *Journal of Sociolinguistics, 5*(1), 50-72.

Siti Aeisha & Hajar A. R. (2014). Corpus research in Malaysia: A bibliographic analysis. *Kajian Malaysia, 32*(1), 17–43.

Tan, S. I. (2013). *Malaysian English: Language contact and change*. Peter Lang: Switzerland