Gema Draft

by Nor Aishah

Submission date: 26-Sep-2022 11:09PM (UTC-0500)
Submission ID: 1910126581
File name: Gema_Draft.docx (340.08K)
Word count: 7684
Character count: 43256

A Corpus-Driven Analysis of Lexical Variation and Change in Malaysian Twitter

ABSTRACT

American English is classified as a hyper-central language; it serves as the hub for global English in Mair's theory of The World System of Englishes. Social media platforms such as Twitter is constantly transforming the usage of lexical items among global Internet users. Interestingly, despite adopting British English in the national education system, American English is gaining prominence among Malaysians due to the widespread dissemination of American English through the media. In spite of the magnitude of American English as a global language, there is a dearth of research on how American English is affecting other varieties of English, especially Malaysian English. There is a need to examine the role of American English in leading global language variation and change. Thus, this study proposes how the influence of American English on Malaysian English, in terms of lexical items on Twitter can be investigated. Using a corpus-driven analysis, this paper will demonstrate how two emerging American lexical items *lit* and *on fleek* can be investigated in terms of its trend of frequency and patterns usage in Malaysian Twitter through concordance and cluster analysis. Results from such a study may be able to reveal the extent to which Malaysian English is influenced by American English in terms of lexical units on Twitter, shedding light on the global transformation of the English language.

Keywords: Lexical variation; lexical change; language variation; language change; Twitter

INTRODUCTION

Language variation and change manifests through phonetic, lexical, and semantic changes. The terms *lexis, lexical items,* and *lexical units* refer to words that can function alone and are oftentimes used interchangeably to signify vocabulary (Lewis et al., 1997; Caro & Mendinueta, 2017). Lexical items are basic units of meaning. Lexical items merge with a set of rules governing the language to produce larger and more complex units such as phrases and sentences.

The recent decade has witnessed a revolutionary change in the English language. For eons, English has been making its way around the world, primarily through historical outreach. The initial use of English was through traditional ways such as speech, sign, and writing. Digital communication accelerated the progress, and still continues to do so. English as the dominant language on the Internet began in the 1990s (Crystal, 2003). The surge in the use of digital technologies over the years with digital transformation technologies such as Cloud, Internet-of-Things, and Artificial Intelligence has brought the ubiquitous use of the Internet, especially social media to the forefront. With the advent of modern technology as a medium of communication, researchers are now concerned with how technologies are moving into new areas in the studies of language. Research has revealed that social media communication leads toward language development (Schmied, 2012), as well as contributing to the development and shift of language practices and linguistic repertoire (Lantz-Andersson, 2018). This has been proven in Sharma (2012) which investigated English language features of Nepal college students and found that their identities were influenced by the introduction of a global social media application, Facebook.

Due to the widespread use of English online, language is undergoing changes by the day. Phonological, morphological, semantic, and pragmatic aspects of the language are changing rapidly. As a result, the use of lexical items especially on social media is also transforming. For instance, the lexical item *google* which derives from the search engine Google came into prominence in 2012, has since then been considered as a verb (Glance, 2015). Another example is the lexis *selfie*, which was declared by the Oxford Dictionaries in 2013 as the Word of the Year due to its phenomenal trend of usage. These are regarded as lexical innovations. Within the strands of English language studies, research on lexical units has expanded in various areas, especially in terms of features (Coats, 2016), analysis and borrowings (List, 2019), emergence and innovations (Huang 2015; Grieve et al. 2017, 2018, 2019), and patterns and variations (Trye et. al, 2020).

One of the online platforms which have propelled language variation and change is Twitter. Founded in 2006, Twitter is a microblogging site which connects users through blogging and instant messages called *tweets*. The use of Twitter has proliferated over the years, with 206 million users worldwide towards the end of 2021 (Brian D, 2022). Twitter has changed and is still changing the social landscape through online communication. There is a wealth of data on Twitter with regard to language studies, and linguists have taken advantage of this by investigating Twitter in language learning (Rosell-Aguilar, 2018), literacy practices (Gleason, 2018), sentiment analysis (Taboada, 2016) discourse analysis (Al-Ghamdi & Albawardi, 2020), and political communication (Santoso, Utari & Kartono, 2020). Recent studies have begun using Twitter in exploring language change and variation, particularly in understanding lexis on Twitter (Huang et. al, 2016; Grieve et al., 2017; Grieve et al, 2018; Grieve et. al, 2019; Würschinger, 2021).

A language reaches global status when it is recognized in every country, and such is the case of English. The global spread of English has resulted in numerous varieties of English over the globe, including Malaysian English. As Baskaran (2005, p. 18) puts it, "after almost two centuries of nurturing and over four decades of nursing, the English language in Malaysia has developed to become a typical progeny of New Englishes: a distinct variety in its own right". The historical context and the language planning policies of the country has enabled the Malaysian variety of English to be an established variety with stylistic layers as well as distinguishing characteristics at the basilectal, mesolectal, and acrolectal levels (Bolton, Botha, & Kirkpatrick, 2020). The essence of the Malaysian variety of English is to provide a sense of identity and to build rapport through its phonology, lexical items, and syntactic structures that are rooted in the Malaysian form (Thirusanku & Yunus, 2012). Many Malaysians now use English as a primary means of communication (Kashinathan & Abdul Aziz, 2021) and traits taken from other languages have been ingrained in the Malaysian variety of English.

The development of the Malaysian variety of English is influenced by historical, linguistic, cultural, and political aspects (Hashim, 2020). More recently, the variety seems to have evolved, especially on social media (Rusli et. al, Hamat & Hassan 2019; Yunus, Zakaria, & Suliman, 2019). New findings have revealed the emergence and innovations of jargons, memes, and slangs, especially among adolescents, who remain as the primary users of social media. With digital communication playing such a pivotal role in language use and development as well as contributing to linguistic identities, ideally, the Malaysian variety of English should be understood by those who speak this variety of English. However, in reality, some features in the Malaysian English are unintelligible to some people, especially with the recent linguistic innovations in the variety because Malaysian English has emerged as a unique variety of English in its lexical usage not only through the infusion of local traits and characteristics (Bolton, Botha & Kirkpatrick, 2020), but also through the impact of global influences and trends (Moody, 2020). While there are numerous studies which have described lexical features in the Malaysian variety of English in terms of localizations (Baskaran, 2005; Baskaran, 2008; Hashim, 2020), there is a dearth of studies which examine the emergence of lexical items in the Malaysian variety of English which are in practice due to the impact of globalisation, particularly on social media.

The English language as well as social media have risen as global forces due to globalisation. This is evident on Twitter, where people connect beyond borders using English. Twitter as a global social networking application is capable of bringing about changes and variations not only to standard English but also to other varieties of English in the world. Thus, since there has been an influx of new global words making their way into the Malaysian variety of English, in which most of them are commonly used on social media, this study is proposed to investigate lexical variation and change in the Malaysian English, particularly on Twitter, to illuminate its distinctiveness as a variety of English that requires linguistic explanation.

LEXICAL VARIATION AND CHANGE

A language's stock of lexical units (lexicon) undergoes changes constantly, whereby there are occurrences of change in form and meaning. Traditional research on lexical variation has distinguished four main types of lexical variations which are semasiological variation, onomasiological variation, contextual variation, and formal variation (Geeraerts, 1994; Miller, 2014). Semasiological variation occurs when a lexical item refers to different types of referents, for instance, *pants* are synonymous with *trousers* and *underpants*. Onomasiological variation happens when a referent or category of referent is named using a variety of conceptually distinct lexical categories. For example, a particular pair of pants can be categorised as a member of pants/trousers category or with a different subordinate category that is jeans. Formal variation occurs when a specific referent or type of referent is named using a variety of lexical items, regardless of whether they reflect conceptually distinct categories or not. For instance, there are a variety of lexical items referring to pants: slacks, trousers, or jeans. Contextual variation refers to the circumstance in which the aforementioned variational phenomena may be correlated with contextual elements such as the formality of the speech situation or the geographical and socioeconomic qualities of the participants in the communicative interaction. For example, the word *pants* is used in informal British English while *trousers* is common in formal British English.

Commented [KS1]: Is there a clearer example for this? This example is quite confusing as pants and trousers can mean the same thing in BE and AE...

Commented [n2R1]: Dear Dr, I am following past studies here and this difference has been used in other studies such as (Miller, 2014).

Semasiological variation and onomasiological variation are the main distinctions that have been used consistently in the study of lexical items. The expansion of the semasiology-onomasiology pair has brought forward processes that affect the changes of meaning (semantic change) and word formation processes in literature. Different scholars have different classifications of word formation processes. A prominent taxonomy is by Yule (1985), which categorised the following under word formation process: coinage, borrowing, compounding, blending, clipping, backformation, conversion, acronym, derivation, prefix, suffix, and multiple processes. Primary word formation processes include inflection, derivation, compounding, clipping, borrowing, blending, truncation, ellipsis, formative extraction, and acronyms (Geeraerts, 2010 & Miller, 2014). Miller (2014) on the other hand, uses the term *lexicogenesis* to encompass word formation processes.

Past studies have shown that lexical change takes place in different ways, one of which is the replacement of certain lexical units with existing words in the language over a certain period of time. For instance, the old English *rood,* which used to represent the cross or crucifix symbolising the cross on which Jesus Christ died, has been replaced by the word *cross* over the years (Miller, 2014). Lexical change is also brought forward by the emergence of new words. This is known as neologism. Neologism is defined by as "newly coined lexical units or existing lexical units that acquire a new sense (Newmark, 1988, p. 140 and are not yet included in general dictionaries (Algeo, 1991). Cook (2010) elaborated that there are two types of neologism, the first type is the combination of words to form novel words, for instance *webisode*, a combination of *web* and *episode*. The second type of neologism is the existing word forms that produce new meanings for example, using email as verb instead of a noun.

After a certain period of time, some neologisms become part of a language's standard and are recorded in the dictionaries. This is known as institutionalisation. Institutionalisation happens when new words conform to a language's norm to become established words and are recognized by a speech community (Bauer, 1988). Different geographical, social, stylistic, and other variants of a language influence institutionalisation. For example, the term *snowman* is not a nameworthy category in African societies but is significant in other societies (Lipka, Handl & Falkner, 1994). Institutionalisation is a gradual phenomenon which provides proper recognition to neologisms, but despite the ability of new words being included in dictionaries, even fully institutionalised terms can become obsolete and disappear. Put simply, novel, fully institutionalised words have the possibilities to either become a permanent part of the language or fall into disuse over time.

New word forms are modelled on prior knowledge, creativity, and imagination (Miller, 2014). Creativity of individuals what yields neologisms across domains and genres. For instance, the political issue of the withdrawal of the United Kingdom from the European Union, famously known as *Brexit* has been in use since 2012 and is now utilized globally (Fontaine, 2017; Jeffries & McIntyre, 2018). A blend of *British* and *exit*, this phenomenon has generated a plethora of new words such as *Brexitology, regrexit, breferendum, Brexitsphere, brexpats, Breturn, Brexitology and brexiteer* through the process of word blending from source words which are Britain, British, and Brexit (Lalić-Krstin & Silaški, 2018). This is evident where language users are capable of exerting their creativity with regard to a socio-political context. Another example is the recent COVID-19 pandemic which has also resulted in the emergence of new words. Research has revealed new acronyms and abbreviations (WFH for work-from-home) as well as the increase in less common expressions before 2020 (self-isolation/self-isolate, physical distancing/social distancing etc). Due to the adverse impact of the pandemic towards the global economy, words such as *furlough* and *layoff* also came into prominence (Asif et al., 2021).

Language research has benefited tremendously from the existence of corpus in observing language change. Corpus is a collection of machine-readable texts, selected consciously according to specific criteria, and serves as linguistic data (Sinclair & Sinclair, 1991). Corpora, plural of corpus are important to allow linguists to make objective statements based on evidence from existing language data instead of constructing subjective statements. Corpora also enable research on language variants from specific periods of time, for instance, language research on dialects or earlier times of a language, feasible now.

Written corpora have been dependent on written samples of newspapers, textbooks, novels, and magazines to derive language data. For instance, newspapers have been used as registers which are represented in popular corpora such as the International Corpus of English, The Brown Corpus, and British National Corpus. Research in recent years has employed digital newspapers as corpora in analysing language variation, particularly in lexical analysis. Lexical contrastive analysis is an area of study which has gained dominance over the years, in which it focuses on the comparison of lexical systems in two or more language systems. One example of research on lexical contrastive analysis is He and Zhou (2015) which investigated the lexical choice between reports on the same accidents between Chinese and American newspapers. The research found that each newspaper has distinct lexical choices that represented their ideological differences and national concerns. Other studies have also explored digital newspapers as corpora, particularly in terms of lexical

features (Endarto, 2020), lexical profiles (Ha, 2021), and lexical borrowing (Kunalan, Mutty & Francis, 2021).

It is important to note that newspapers are only a rough representation of speakers' experiences. This style of work is carefully edited and focuses on a few, specific issues that are not related to everyday social interactions (Brysbaert & New, 2009). They have a register that is considerably different from how people speak naturally and spontaneously. The emergence of world wide web and social media has allowed researchers to have a greater understanding of spontaneous, real-life language change, especially in written communication. Personal emails, chat rooms, online forums, instant text messaging applications such as WhatsApp as well as platforms such as Facebook, Twitter, Blog, and Instagram with their unique appeals have enabled the global speech community to connect and express themselves via words. This is referred to either as digital networked writing (Androutsopoulos, 2011), computer-mediated communication (Romiszowski & Mason, 2013), or computer-mediated discourse (Herring & Androutsopoulos, 2015. According to Arrizabalaga (2021), past literature has delineated the following as the distinguishing features of language use on the Internet: (i) replete with acronyms, emoticons, emojis, contractions, repeated letters, capital letters with connotative meanings, inventive use of punctuation marks, unusual spellings, and self-corrections; (ii) spurred by creativity and innovation, resulting in new word forms derived from different word processes, and (iii) rife with omissions, incomplete clauses, and informal expressions. The distinct use of language online has contributed to language variation and change.

This social media's linguistic transformation (Tankosić & Dovchin, 2021) is propelled by language users online. Graddol (2000, p. 51) stated that the Internet "has given the shift of control to ordinary users". The power of traditional media such as printing and broadcasting which were once the gatekeepers to promote standard language have shifted to the Internet. As a result, the Internet, especially social media, is "contributing to the fluidity and promotion of vernacular, or in-group, language" (Battarcharjee 2009, p. 49). With such extensive and collective power belonging to social media users worldwide, the role of propelling linguistic transformation now lies with them. Language change in social media is now driven by a global network of users. The world is now witnessing the birth of global lexical items, shared by global citizens.

Giddens (1991) defines globalisation as the strengthening of global social relations that connect far-flung locations in such a way that local events are impacted by events taking place thousands of miles away, and vice versa. Globalisation has enabled linguistics to be observed beyond traditional constructs and barriers, addressing burgeoning ideas which are defining the society. Blommaert (2010) referred to this as 'sociolinguistics of globalization'. The goal of sociolinguistics of globalisation is to connect ideas that go beyond a stratified, unidirectional perspective of the language, by understanding "trans-contextual networks, flows, and movements" (Blommaert, 2010, p.1). The English language is continually flowing beyond traditional geospace, causing a shift in English practices and introducing new conventions. Tsang and Hinrichs (2020) expanded on the notion of 'language mobility', in which the English language is considered to be traversing all around the world through various processes. Traditional research has equated mobility with migration and contact, but over the years, English has evolved to accommodate social processes such as globalisation and concepts such as transnational flows. Language mobility is now affecting general aspects of language use and appropriation as well as personal aspects like identity of language users.

The mobility of English language today is no longer constricted to geospace; English flows through digital spaces too. Our increasingly global interests and outlook have been in large part facilitated by the advent of the internet and the subsequent proliferation of technologies and platforms for sharing and disseminating information. According to Androutsopoulos (2011), global circulations are conceptualised at two main levels. The first takes place when new genres or discourse patterns emerge on a larger scale, such as in news reporting, businesses, or popular music. The second level is when linguistic features, particularly lexical items, spread across dialects or languages.

Lexical items used globally affect collective and individual lexical repertoires. As the global lexicons expand, collective and individual lexical repertoires also change - either increasing or decreasing in size. Androutsopoulos (2014) investigated language practices on Facebook and the impact it had on individual linguistic repertoires, and found a connection between the two. Tankosić and Dovchin (2021) investigated the impact of social media towards peripheral countries such as Bosnian, Serbia, Mongolia and found that peripheral languages adopt relocalisation. Relocalisation is one of the impacts of globalisation, in which lexical items and discourse markers from English are borrowed and re-adapted into local alphabetic, orthographic, syntactic, and grammatical systems to the extent where the original speakers are unable to understand them (Androutsopoulos, 2011).

The most important part of global spread is the proliferation of varieties of English, an unending compendium of regional, national, subnational, ethnic, and pidgin and creole varieties (Canagarajah, 2016). Pennycook (2007) states that "Languages will flow and change around us, new combinations of languages and cultures will be put together, texts will be sampled and mixed in ever new juxtapositions" (p. 158). This statement aligns with the concept of linguascapes, which Dovchin (2018) defines as "the transnational flows of linguistic resources circulating across the current world of flows, making meanings in contact with other various spatiotemporal scapes interacting with one another, and affecting the particular speakers' linguistic practices in varied ways" (p. 35).

Pennycook (2007) affirms that there is a necessity to evaluate the spread and use of English around the world while taking into account the various local contexts in which it is used, such as history and politics, the current linguascape, language ideologies, economy, and infrastructure. Pennycook (2007) expanded:

> "At the very least, we need to understand how English is involved in global flows of culture and knowledge, how English is used and appropriated by users of English around the world, how English colludes with multiple domains of globalization, from popular culture to unpopular politics, from international capital to local transaction, from ostensible diplomacy to purported peace-keeping, from religious proselytizing to secular resistance" (p. 19).

Tseng and Hinrichs (2020) concurred that future research could address underexplored issues surrounding English language mobility and contact, such as English in digital media and transnational networks. New global electronic discourses have emerged from online chat, instant messaging, weblogs, podcasts, and mobile apps as a result of the impact of digital media on language users on a daily basis, paving new research directions in global Englishes that have not yet been thoroughly explored (Martin, 2019), and such research is necessary because "it will help determine whether the media consistently and

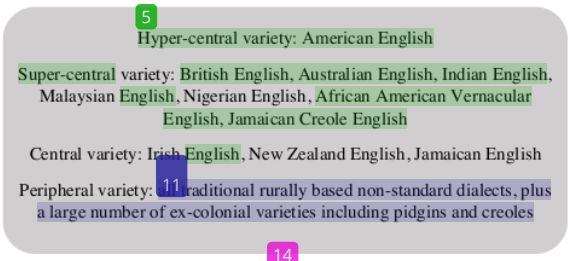
accurately reflect the "pluricentricity" of English or, on the contrary, largely misrepresent both linguistic and sociocultural reality" (p. 607).

Despite the fact that English is widely acknowledged to be prevalent in global computer-mediated communication, there is a death of research in studying the lexical units or grammar of English on social media (Coats, 2016). Clearly, there is a need to study global lexical items which have emerged in different varieties of English. However, before investigating lexical variation and change online, it is important to first dissect the global role of the English language, which will be covered in the next section.

THE WORLD SYSTEM OF ENGLISHES

Over the past decades, English has come to the fore as one of the widely used languages in the world. The study of English as a global language could be traced back to Kachru (1985) when he developed the theory of World Englishes. The term *Englishes* is used to signify the different dialects and variations of the English language and "to capture distinct identities of different Englishes and to critically examine the implications of such identities in cross-cultural communication" (Kachru, 1992, p.2). Kachru's theory of World Englishes has received appreciation for recognizing the different varieties of English and the paradigm has emphasised the necessity of seeing English from various perspectives. However, despite its significance, this framework has also been disputed in literature mainly for being constrained by the geography and history of nations (Pennycook, 2007).

Kachru's World Englishes have paved the way for the emergence of new theoretical frameworks which accounts for the untenable aspects in the theory, such as social and linguistic factors. One such theory is The World System of Englishes. Proposed by Mair



(2013), this theory encompass and explain all the varieties of English around the world. There are four levels in this model, which are hyper-central variety, super-central varieties, central varieties and peripheral varieties.

Figure 1: Mair's Theory of The World System of English (2013)

Language users of peripheral varieties should be well-versed in and at least partially active in a number of other languages, especially those spoken by individuals at the top of the social order. For speakers from central regions, it is common practice to adapt the pertinent super-central materials. For instance, Irish English users would adopt British English as their language references. This theory regards American English as a hyper-central language, functioning as the hub for global English.

The theory of The World System of Englishes postulated that American English has established itself as the most-central variety, having the greatest potential to impact other varieties of English. This is largely due to the global spread of American English through media_- in films, newspapers, advertisements, and broadcasting (Crystal, 2001). The eminence is now added with the presence of social media. Interestingly, while British English continues to be the dominant yardstick in the majority of English education systems around the world, American English is gaining traction globally, particularly in non-native countries due to the large number of native speakers and dominance over popular culture. Igeo (2006), followed by Gilquin (2018), affirmed that the dichotomy between British and American English is propelled by three main factors: (1) education system, (2) number of native speakers, and (3) popular culture.

American popular culture is a "global juggernaut" (Crothers, 2021, p. 234), dominating movies, music, television and now digital communication. These sources of entertainment and communication are channelled and produced in American English, which results in the ubiquity of American English. The population worldwide is exposed to American English on a daily basis. The presence of the global digital environment in recent years is blurring the geographic lines, causing gradual transformation towards other varieties of English. American English, being increasingly abundant on social media, especially on American-owned platforms such as Twitter is driving global language variation and change.

In Mair's theory of The World System of Englishes, the key indicator of the influence of American English is when lexical units from American English spread into other varieties of English. This is known as lexical Americanisms. Mair elucidated that lexical traffic or the direction of flow of lexical units will occur "downwards" instead of "upwards", which means lexical borrowings are expected to take place according to the hierarchical levels. For example, lexical units from American English (hyper-central variety) are more likely to spread into Malaysian English which is lower in hierarchy (super-central variety).

The Malaysian variety of English has witnessed a wide expanse of language variation, firstly due to the impact of colonialism and now because of globalisation. Pillai (2017) asserted that as Malaysian English "began to be used by a wider range of speakers and in expanding contexts, distinct linguistic and socio-pragmatic features began to emerge as these users adapt English to suit local tongues, norms and nuances" (p.19). Thus, based on the theory of The World System of Englishes, this study is proposed to investigate how the global widespread and pervasiveness of American English is influencing Malaysian English on Twitter, especially in terms of lexical units. This will be covered in detail in the next section.

LEXICAL VARIATION AND CHANGE ON TWITTER

Social media have expanded the frequency and speed at which we could communicate; thus, language change is now more rapid than ever. Twitter as a global platform is widely utilized, and because of this the language used on Twitter is now easier to observe, disseminate, and acquire. When a public tweet is published on Twitter, it becomes available to everyone in the globe, whether through liking, replying, retweeting, or forwarding. Twitter, with its instant, real-time features, allows for the propagation of information, knowledge, communication, and ideas to take place easily which then transforms the English language with new developments. Due to the novelty of this area of research, relevant studies on lexical variation and change on Twitter are elaborated in detail in this section.

With the upsurge in social media use, especially Twitter, linguists are now able to have access to a large amount of linguistic data, which has and could fuel more research in language variation and change. The massive amount of data which can be retrieved online, otherwise known as Big Data, allows access to the collections of information composed of natural, publicly available data. Big Data, combined with computational methods, could bring numerous new, interesting discoveries on different subjects to the forefront. In studying lexical variation, Big Data "increases scope and enhances granularity of study, allowing rare and intuitively inaccessible features to be glimpsed" (Renouf, 2018, p.27). One of the key features in Twitter which enables research on lexical units is the spatial and temporal continuity it offers (Huang et al., 2016). The development of mobile systems with Global Positioning System (GPS) has allowed tweets on Twitter to have both time annotations and spatial information. Tweets on Twitter are stored from its inception until now; therefore, it consists of a large amount of historical and real-time data.

Early studies used to download tweets manually but tweets can now be harnessed through Advanced Search feature and the Twitter API Developer platform, which enables both historical and recent tweets to be mined into a corpus. This is known as corpus compilation. Corpus compilation is the act of "designing a corpus, collecting texts, encoding the corpus, assembling and storing the relevant metadata, marking up the texts where necessary and possibly adding linguistic annotation" (McEnery & Hardie, 2012, p.241). Researchers have made use of corpus compilation using web sources in analysing lexical variation and change, because corpus analysis is the only methodology for quantitative assessments of diachronic change, and is the most popular tool for examining synchronic variation (Krug, Schlüter & Schluter, 2013). For instance, the Corpus of Global Web-based English (GloWbE) and The NOW (News on the Web) Corpus extract language data from online sources which enables both synchronic and diachronic investigation on lexical items. Another web source which has benefitted the study on lexical variation and change is the Google Books archive, for example Petersen et al. (2012) examined the dynamic features of

words in English, Spanish, and Hebrew words recorded between 1800 and 2008 using Google NGram.

Since tweets are geo-tagged and time-stamped, they allow researchers to investigate lexical variation and conduct geographical analysis or the mapping of linguistic characteristics using data on Twitter. To understand research on regional patterns, we need to look back at the inception of this research area. Labov (1963) pioneered the research on regional lexical variation, particularly in phonological variation. Early research in dialect variation including Labov's (1963, 1966) has traditionally employed methods such as fieldworks by entering the community and having one-on-one interviews with informants to examine regional variations. Labov's (1966) most widely discussed study investigated sound change in New York City in the way New Yorkers use the phoneme /r/, and his subsequent research has contributed profoundly to the field of variationist sociolinguistics. The contribution of traditional data is not without its limitations. The main problem in collecting language data using fieldwork and interviews is the Observer's Paradox, the lengthy period of time required to collect data.

According to Grieve (2019), there are several advantages of using Twitter corpora, especially in conducting research for regional lexical variation. The first advantage is that corpora is easier to build using Twitter data than gathering data using fieldworks and surveys. To illustrate, before a fieldwork can take place, it is important to determine the exact linguistic variables that the researcher intends to investigate (Feagin, 2013). On the other hand, Twitter corpora enable open-ended analysis of a much broader variety of linguistic features. Thirdly, Twitter corpora could eliminate Observer's Paradox (informants modifying their speech due to the presence of an observer) and examine language in a natural state. Another advantage of Twitter corpora is that it improves the resolution of dialect maps, in which it allows more informants to be sampled in more places.

In the recent years, Twitter data has been utilized to study lexical dialect variation (Eisenstein et al. 2014; Doyle, 2014; Jones, 2015) and map regional dialects. Huang et. al (2016) investigated regional variation of American English using lexical alternations i.e., variations of a particular word with the exact or similar meaning, for instance mom/mother. By using one year of Twitter data, they mapped lexical alternations produced by American Twitter users according to the counties in the United States through principal component analysis and regionalization methods. The research revealed unique linguistic characteristics according to the regions. Some alternations were found to be similar with certain regions while other alternations were different. People in the Northeast, for instance, preferred *bag* over *sack* and this preference is much less pronounced in the South. The word *clearly* is preferred in the East whereas *obviously* is used more frequently in the West.

To assess the generalizability of Twitter data in examining regional variations, Grieve et al. (2019) investigated lexical dialect variation, particularly lexical alternations in British Twitter. This study compared 1.8-billion-word corpus of geolocated UK Twitter data with traditional survey data from BBC Voices Project and found broad alignments between the two datasets. The findings confirmed the reliability and effectiveness of Twitter as a resource to study dialect patterns. Regional dialect mapping using Twitter data is found to be propitious compared to survey data because of the greater accuracy in the identification of regional patterns.

With each passing day, there are more new word forms making their way into the English language and spreading among English users. This is known as lexical emergence. There is a dearth of research on lexical emergence because "linguists have not had access to sufficient amounts of language data with the necessary temporal resolution to track the spread of emerging word forms" (Grieve et. al, 2017, p.102). Thus, to overcome this, Grieve et. al (2017) introduced a methodology to investigate emerging lexical units in American Twitter. Through calculations of relative frequency, measurement using Spearman correlation coefficient as well as concordance analysis of one-year Twitter data, 29 emerging word forms in American Twitter were revealed. Some of the new word forms include *rekt, lit, faved, on fleek, tooka, mutuals* and so on.

Following the procedures introduced in Grieve et al. (2017) to investigate lexical emergence, Grieve et. al (2019) studied lexical innovations in American Twitter by mapping the origin and diffusion of the lexical units. The research found urban regions which are rich in culture to be the main hotspots of lexical innovations. These regions were California, Atlanta, New York City, Washington D.C., and New Orleans. Some of these hubs of lexical innovations were also largely dominated by African American English, which is the primary source for new emerging forms in American English. This is in line with a prominent study by Pennycook (2007) which asserted that African American culture is a powerful force in global change, especially as a tool for redefining local identities all around the world.

It is now apparent that textual data harnessed from Twitter allows for further linguistic analysis. There are several key linguistic methods which has been used in the aforementioned research to investigate lexical items from Twitter, including distribution and trends of lexical items. Frequency distribution is frequently studied in corpus linguistics, in which the occurrence of particular lexical items in a corpus is examined. The results obtained from frequency distribution can then be inspected to observe the trends of usage of a particular lexical item in a corpus. For example, Awal et. al (2021) investigated the trend of frequency of the Islamic terms *halal* and *haram* in the Malaysian Hansard Corpus and observed its patterns across 13 parliament sessions from the year 1959 until 2018. Directly relevant to this paper is the recent research by Giorgi et al. (2022) which investigated the frequency distribution of *BlackLivesMatter*- a social call against racism which has garnered attention over the years, especially on social media. The study examined the distribution of tweets on *BlackLivesMatter* across the United States for three 3-year periods: 2013 to 2015, 2016 to 2018, and 2019 to 2021.

Google Trends has been utilized as a platform to complement results from frequency distribution (Grieve, 2017). Google Trends is used to trace specific search words or phrases either synchronically or diachronically. For example, by searching the term *selfie,* the platform reveals numerical and graphical data regarding *selfie* which can be tailored to researchers' desired region, country, categories, and duration. The numerical breakdown of the usage of *selfie* can be downloaded for further analysis. Google Trends also allows the generation of the geographic maps for the term *selfie* to illustrate where the word is most prevalent.

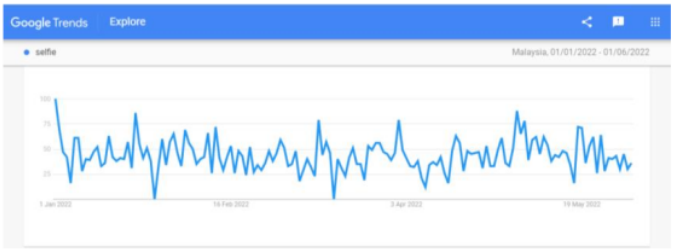


FIGURE 1. Google Trends result for the term *selfie* in Malaysia for the first six months of 2022. Data source: Google Trends (https://www.google.com/trends).

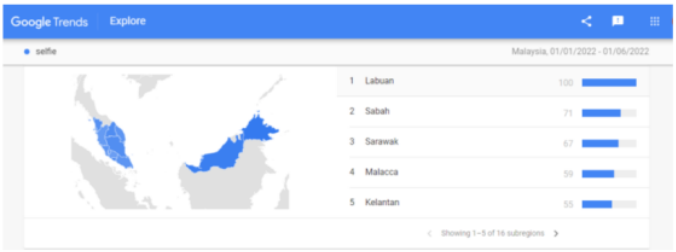


FIGURE 2. Geographic maps generated in Google Trends for the term *selfie* in Malaysia for the first six months of 2022 according to the subregions in the country. Data source: Google Trends (https://www.google.com/trends).

Language use on Twitter is spontaneous and recent. We now witness new word forms which are foreign to us making their way into the English language and spreading among Twitter users. To understand the meaning and context of usage of these new words, researchers have made use of available tools, such as concordancer in corpus analysis toolkits. With the aid of concordance, specific words (also known as target item or node word) are displayed alongside context words to reveal how the terms are used in a corpus. For instance, in Taufek et al. (2021), based on the sentiment lexicon revealed in the corpus, concordance was utilised to infer public sentiments towards climate change. To understand new lexical units on Twitter, Grieve et. al (2017) used concordance to understand how each of the novel lexical items were used on Twitter.

Due to the novelty of emerging lexical items, extra steps need to be taken to ensure the accuracy of meanings of the lexical items found online. Therefore, apart from concordance analysis there are additional tools used to further analyse novel word forms. To investigate language change on Twitter, webpages such as Urban Dictionary have been

utilized to track the emergence of new words. Urban Dictionary is a crowd-sourced online dictionary founded in 1999 and has since then been utilized to share the definitions of new word forms available online. The importance of Urban Dictionary in 17mining language change on Twitter is two-fold. Firstly, this user-generated site helps to track the earliest occurrence of words and phrases and second, Urban Dictionary enables researchers to verify the definition of terms found in Twitter corpora (Grieve, 2017).

Lexical units, despite its ability to stand on its own, are mostly used with other lexical units to form multi-word combinations or recurrent word sequences, with phrases consisting of at least two words or more (McEnery & Hardie, 2012). They are used interchangeably in literature using the following terms: phraseological units, n-grams, multi-word lexemes, clusters, prefabricated speech or prefabs, fixed expressions, lexical bundles, set phrases, phrasemes, formulaic sequences (Byrd & Coxhead, 2010; Fiedler, 2018). Two of the most-used terms are lexical bundles and phraseological units. Lexical bundles are mostly associated with academic register– in investigating research articles (Varghaei, Branch & Khodadadi, 2022), dissertations (Narkprom & Phoocharoensil, 2022), and textbooks (Hussain, Zahra & Abbas, 2021). According to Khayrullina and Fatkullina (2021, p. 273), phraseological units can be divided into grammatical and semantic features. Grammatical structures can be divided into communicative phrases which form sentences (*Mothers, they are like that!*), phraseological phrases (*to wither on the vine*), and unit as word forms (*So what?*). Semantic-wise, phraseological units can be classified into thematic groups according to the socio-cultural functions of different language groups (slang, socio-political, medicine etc). To illustrate an example, *on fleek* is commonly used in American Twitter as a phraseological unit- *eyebrows on fleek* (Grieve et al., 2017).

THE CASE OF *LIT* AND *ON FLEEK* IN MALAYSIAN TWITTER

The proposed study will investigate two lexical items - *lit* and *on fleek* - which emerged in American Twitter since 2013 (Grieve, 2017). These lexical items will be investigated in Malaysian Twitter from the year 2013 until 2021 in terms of frequency distribution, the usage of these lexical items as well 29he phraseological units formed from these lexical items. The results from this study will shed some light on the extent to which Malaysian English is influenced by American English, particularly on Twitter. As elaborated earlier, the findings will be discussed according to the theory of World System of Englishes, as well as relevant insights on the impact of globalisation on lexical change and variation via digital communication.

Trends of frequency enable researchers to gain a comprehensive understanding of how lexical items are distributed in a corpus and draw conclusions from them. In this study, the terms *lit* and *on fleek* will be investigated for its distribution of usage from the year 2013 until 2021, since these word forms emerged in American Twitter from the year 2013 onwards (Grieve et. al, 2017). Frequency will be used to ob8rve the progression of these terms over the years, and to achieve this, absolute frequ8cy (the number of times a lexical item occurs in the corpus) as w22 as relative frequency (the number of times a lexical item occurs in the corpus in relation to the total number of words i.e., tokens in the corpus) will be calculated. After the calculations are performed, the data will then be visualized using a line chart. This methodology would reveal the trends of usage of these lexical items over the years in Malaysian Twitter.

In this research, apart from the aforementioned frequency distribution analysis, Google Trends will also be utilized to further understand the earliest occurrence of *lit* and *on*

fleek in Malaysia. To attain this, *lit* and *on fleek* will serve as the search terms, and the following search criteria will be customised on Google Trends: (1) location set to Malaysia; (2) time range from the year 2013 until 2021; (3) categories set to all categories for a comprehensive result; (4) web search is selected. This would reveal the distribution of these lexical items on Google during the specified time frame. By performing the aforementioned methodology, Google Trends will concurrently reveal the distribution of these lexical items in all the sub-regions in Malaysia in a geographical map. Sub-regions here refer to the states and federal territories in the country, i.e., the 13 states and 3 federal territories in Malaysia.

As discussed earlier, concordance helps researchers understand the ways lexical items are actually employed in a corpus, and this study will also make use of the concordance. Tweets with the lexical units *lit* and *on fleek* will be analysed using the concordancer in AntConc version 4.1.1 to analyse the usage of these word forms in Malaysian Twitter. Additionally, Urban Dictionary will be used to verify and compare the usage of lexical items *lit* and *on fleek* in Malaysian Twitter with the meanings in American English. This would demonstrate whether Twitter users in Malaysia use these lexical terms similarly to American Twitter users and are referring to the same meanings when they do.

The lexical items *lit* and *on fleek* are not used in isolation. As elaborated earlier, *on fleek* is oftentimes accompanied by *eyebrows* to form *eyebrows on fleek*. Due to this, this study will identify the common phraseological units of *lit* and *on fleek* in Malaysian Twitter and compare its usage with phraseological units in American English, which can be found on Urban Dictionary. To identify the use of these phraseological units, the n-gram feature in corpus tools will be utilized. AntConc version 4.1.1 allows for the common patterns of word sequences and its frequency to be revealed through its n-gram/cluster feature. For instance, by keying in the term *on fleek*, with the minimum phraseological units accepted to be two or more words as per Harris (2006) and McEnery (2011), the corpus tool would reveal the existence of one of the common phrases in American English that is *eyebrows on fleek* in Malaysian Twitter. The phraseological units found will then be classified according to Fiedler's (2007) classification of phraseological units.

It is important to note that the English language use in Malaysia is rife with code-switching, whereby language users alternate between two or more languages at once. Being a multilingual nation, there are sequences of words adopted from other local languages in Malaysia included while practicing the English language, and this phenomenon is apparent on social media platforms such as Twitter. This study strictly focuses on the usage of *lit* and *on fleek* in the English language in Malaysian Twitter, with tweets being fully in English language. This means that if Malaysian Twitter users use *your kening is on fleek*, with *kening* referring to *eyebrows*, replacing the common phraseological unit *eyebrows on fleek*, it would not be included in the main analysis. Nevertheless, considering the uniqueness of code-switching among Malaysians, this aspect would still be taken into account in the analysis and discussion to show the localized features in Malaysian English but only as additional findings, and not as the main results of the study.

In summary, this section has propounded the methodology in analysing lexical variation and change in Malaysian Twitter. The procedures which will be taken has been utilized in past research, and this study has systematized the necessary steps to investigate a particular variety of English through a corpus-driven analysis. Some of the pertinent tools and

analysis are frequency distribution, the use of Google Trends, concordance, Urban Dictionary for verifications, and n-gram/cluster feature in AntConc version 4.1.1.

CONCLUSION

The results from such a study can provide a general overview of the emergence of new global words in Malaysian English, to establish the predominant word-formational patterns as well as to observe the usage of these words online. The findings will be analysed and discussed according to the theory of World System of Englishes, as well as the recent theories of globalisation and language mobility in sociolinguistics. Apart from that, the present study could lead to theoretical contributions by illuminating the dynamics of English language variation and change at large.

As this study is driven by the need to understand lexical variation and change in Malaysian English (super-central variety) and comparing it to American English (hyper-central variety) to understand the extent to which Malaysian English is influenced by American English, particularly on Twitter, future studies could potentially expand the proposed study by investigating other varieties of English; i.e., central and peripheral varieties in the theory of World System of English. The advancement of technology in propelling digital communication will most definitely result in more incoming novel lexical items, catering to either specific groups of language users, i.e., youth or English language users at large. The subject of lexical variation and change can still be explored further in the upcoming years, and it is hoped that the purported study has shed some light on this research area.

Gema Draft

ORIGINALITY REPORT

7%	3%	5%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Modeling World Englishes", John Benjamins Publishing Company, 2018 Publication	1%
2	"The Handbook of World Englishes", Wiley, 2019 Publication	<1%
3	Submitted to Universiti Teknologi MARA Student Paper	<1%
4	"Changing English", Walter de Gruyter GmbH, 2017 Publication	<1%
5	"World Englishes", John Benjamins Publishing Company, 2016 Publication	<1%
6	www.ncbi.nlm.nih.gov Internet Source	<1%
7	Submitted to University of Brighton Student Paper	<1%
8	Submitted to University of Canterbury Student Paper	<1%

9	Submitted to University of Queensland Student Paper	<1%
10	Submitted to Kennesaw State University Student Paper	<1%
11	www.mediensprache.net Internet Source	<1%
12	Submitted to The Hong Kong Institute of Education Student Paper	<1%
13	journals.sagepub.com Internet Source	<1%
14	"Ugandan English", John Benjamins Publishing Company, 2016 Publication	<1%
15	uir.unisa.ac.za Internet Source	<1%
16	Submitted to University of Reading Student Paper	<1%
17	"The Handbook of Historical Sociolinguistics", Wiley, 2012 Publication	<1%
18	3lib.net Internet Source	<1%
19	hdl.handle.net Internet Source	<1%

20	"From Data to Evidence in English Language Research", Brill, 2019 Publication	<1 %
21	asunow.asu.edu Internet Source	<1 %
22	silo.pub Internet Source	<1 %
23	research.birmingham.ac.uk Internet Source	<1 %
24	"Africa, South and Southeast Asia", Walter de Gruyter GmbH, 2008 Publication	<1 %
25	"Language, Education and Technology", Springer Science and Business Media LLC, 2017 Publication	<1 %
26	Christian Mair. "The World System of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars", English World-Wide, 2013 Publication	<1 %
27	archive.org Internet Source	<1 %
28	docslib.org Internet Source	<1 %

29	doi.org Internet Source	<1%
30	dokumen.pub Internet Source	<1%
31	umexpert.um.edu.my Internet Source	<1%
32	"Functional Variations in English", Springer Science and Business Media LLC, 2020 Publication	<1%
33	"Modeling world Englishes from the perspective of language contact : Modeling world Englishes from the perspective of language contact", World Englishes, 2016. Publication	<1%
34	link.springer.com Internet Source	<1%
35	openaccess.maltepe.edu.tr Internet Source	<1%
36	studentsrepo.um.edu.my Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

Gema Draft

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16