GEMA Online® Journal of Language Studies
Vol. 26(2), May 2026 http://doi.org/10.17576/gema-2026-2602-02

English Loanwords in Malay: Identifying Patterns in Academic Magazines
(Kata Pinjaman bahasa Inggeris dalam bahasa Melayu: Pengenalpastian Pola dalam Majalah Akademik)

ZURAIDAH MOHD DON [1], GERRY KNOWLES [2] & NOR SHAHILA MANSOR [3]

[1] *UCSI University, Malaysia*
[2] *Independent Researcher, United Kingdom*
[3] *Universiti Malaya, Malaysia*

Received: 27 October 2026 / Accepted: 14 May 2026

ABSTRACT

English loanwords have become an important feature of contemporary Malay, particularly in domains shaped by global knowledge, technology and academic discourse. However, previous studies have often treated loanwords as descriptive lists or as general evidence of language contact, with less attention to their lexical distribution and structural integration in Malay. This study adopts a corpus-based approach to examine English loanwords in Malay academic magazine texts and to identify patterns that reveal their broader linguistic significance. The data comprise approximately two million words of well-formed Malay academic magazine articles, supported by a digital Malay lexicon containing 12,504 lemmas and 49,933 lexical words. Using corpus linguistic procedures, the study distinguishes between lemmas and individual lexical items, identifies loanwords of English origin, and analyses their grammatical classes, morphological behaviour and frequency patterns through a relational database system. The findings show that English loanwords constitute 3,982 lemmas, or 32% of the lexicon, but only 6,794 lexical words, or 14% of the lexical items, indicating that they are numerous but relatively less morphologically productive. Most borrowings are nouns, while directly borrowed verbs are rare. However, Malay morphology forms verbs from borrowed nouns, especially through a certain circumfix. The study also identifies borrowed complex adjectives as a possible emerging subclass in Malay. Overall, the findings suggest that English loanwords not only expand the Malay lexicon but also contribute to ongoing structural innovation in contemporary Malay.

Keywords: Global English; contemporary Malay; loanwords; corpus linguistics; process innovation.

ABSTRAK

Kata pinjaman bahasa Inggeris telah menjadi ciri penting dalam bahasa Melayu kontemporari, khususnya dalam bidang yang dipengaruhi oleh pengetahuan global, teknologi dan wacana akademik. Walau bagaimanapun, kajian terdahulu sering memperlakukan kata pinjaman sebagai senarai deskriptif atau sebagai bukti umum berlakunya kontak bahasa, dengan kurang memberikan perhatian terhadap taburan leksikal dan integrasi strukturnya dalam bahasa Melayu. Kajian ini menggunakan pendekatan berasaskan korpus untuk meneliti kata pinjaman bahasa Inggeris dalam teks majalah akademik berbahasa Melayu serta mengenal pasti pola yang memperlihatkan kepentingan linguistiknya secara lebih luas. Data kajian terdiri daripada kira-kira dua juta patah perkataan daripada artikel majalah akademik berbahasa Melayu yang tersusun dengan baik, disokong oleh leksikon digital bahasa Melayu yang mengandungi 12,504 lema dan 49,933 kata leksikal. Dengan menggunakan kaedah korpus linguistik, kajian ini membezakan antara lema dengan item leksikal individu, mengenal pasti kata pinjaman yang berasal daripada bahasa Inggeris, serta menganalisis kelas tatabahasa, tingkah laku morfologi dan pola kekerapannya melalui sistem pangkalan data hubungan. Dapatan kajian menunjukkan bahasa kata pinjaman bahasa Inggeris merangkumi 3,982 lema, atau 32% daripada leksikon, tetapi hanya 6,794 kata leksikal, atau 14% daripada item

* *Corresponding author: zuraidah@ucsiuniversity.edu.my*

eISSN: 2550-2131
ISSN: 1675-8021

leksikal. Hal ini menunjukkan bahawa kata pinjaman tersebut banyak dari segi bilangan, namun secara relatifnya kurang produktif dari segi morfologi. Sebahagian besar kata pinjaman tersebut terdiri daripada kata nama, manakala kata kerja yang dipinjam secara langsung adalah jarang. Walau bagaimanapun, morfologi bahasa Melayu membentuk kata kerja daripada kata nama pinjaman, khususnya melalui apitan tertentu. Kajian ini turut mengenal pasti kata adjektif kompleks pinjaman sebagai kemungkinan subkelas Baharu yang sedang muncul dalam bahasa Melayu. Secara keseluruhannya, dapatan kajian menunjukkan bahasa kata pinjaman bahasa Inggeris bukan sahaja memperluas leksikon bahasa Melayu, malah turut menyumbang kepada inovasi struktur yang sedang berlaku dalam bahasa Melayu kontemporari.

Kata kunci: Bahasa Inggeris global; bahasa Melayu kontemporari; kata pinjaman; korpus linguistik; proses inovasi

INTRODUCTION

Loanwords have been studied conventionally in the context of languages in contact, when speakers of some language borrow words from some other language which they regularly encounter (Matras, 2009; Thomason, 2021). This is known as "a language-contact phenomenon" (Havumetsä (2022, p. 562). Since at least the twentieth century, English has been the *de facto* language of knowledge and innovation, and accelerating globalisation has brought English into contact with other languages. English has taken on the role of conduit from global culture to languages around the world, and has influenced Malay much as it has influenced other languages. Despite the national desire to preserve the linguistic purity of Malay, especially its vocabulary, Malay like other languages (Durkin, 2014; Portugal & Nonnenmacher, 2024) also borrows lexical terms from foreign languages, especially English (Michaud & Hollenback, 2015). In the globalised world, the sheer volume of needed new words makes it impracticable to coin words separately for different languages, and the realistic solution for Malay, as for many other languages, is to make use of words already formed in English (Platt, 2024).

The traditional outcome of loanword research is a list of loanwords which falls short of representing the impact of loanwords on the recipient language. The research problem is that the role of English as the general source of new words makes it impossible to devise a list of loanwords that can be said to adequately represent the current state of the art. The approach adopted here to fill the gap and solve the problem is to make an objective study of the words that occur in a large authentic, naturally-produced data set. Because this data set is necessarily an arbitrary sample, we have to expect that other samples will contain different sets of words. We cannot therefore claim that our loanword list represents the contemporary Malay language. What is possible at this stage is to look for patterns in the data unlikely to have arisen by chance. These can be said to represent the way English loanwords are used in Malay, and make it substantially possible to replicate the main findings, albeit with different word sets.

The study of a large data set requires the methods of corpus linguistics (Biber & Reppen, 2015; Bond, 2025). This paper is a spinoff from the MaLex project (2020), which aims to make a digital model of Malay. It uses computational means to create a language description which simulates the intuitive knowledge of Malay possessed by L1 speakers of the language. The output consists of declarative knowledge of Malay in the form of tables and code which can in principle be exported and used for other projects, including this one. MaLex in this way provides an appropriate environment for the study of English loanwords.

This introduction locates the study of English loanwords in Malay in the context of worldwide borrowing from English. The next section discusses publications relevant to the corpus-based methodology used here. Section 3 outlines the data and methods, which go beyond the

conventional listing of loanwords, and enables the measurement of the penetration of words into the Malay language system. Section 4 reports the findings, and this is followed by a discussion and a separate conclusion.

CURRENT ISSUES IN THE LITERATURE

Loanwords constitute a significant component of the lexical inventories of many languages (see e.g. Haspelmath and Tadmor, 2009; Durkin, 2014). This prominence is however not reflected in the amount of scholarly attention devoted to the topic, leaving some issues inadequately explored (Zenner et al. 2019). The review is accordingly concerned not with individual publications on loan words, but with current issues in the study of loanwords which have affected the design of the project on which this paper is based. This constraint makes it impossible to give more than a mention to historical work including Morrow (2020), who uses the Corpus of Historical American English (COHA) to trace the changing frequency of Japanese words over time, and Oh & Son (2023), who trace increasing numbers of English loanwords in Korean, especially after 1990.

BASIC WORDS AND LEXICAL SETS

It has long been normal practice in the study of loanwords to compile and study lists of basic words. In the case of English, the first word lists were probably produced by Mulcaster (1582), and the modern study of loanwords began in philology closely related to etymology (e.g. Trench, 1851), using lists of basic words. More recent lists, e.g. Fries & Traver (1960) have been devised for educational purposes. In this long tradition, words and loanwords are treated as matters of common sense, and so assumed to be generally understood. Close inspection, however, reveals that these terms are not precisely defined, and vary in meaning from one situation to another. This leaves a gap which this paper aims to fill, especially as the use of corpus-based methods brings with it in any case the need for greater precision. Our first task here is to make clear what is meant by words and loanwords.

A problem of definition involving loanwords is illustrated by the work of de Heer et al (2023), who report recent research on Uralic languages. They do not define their loanwords or related terms such as basic vocabulary (p. 54), and base their analysis on "a dataset of Uralic basic vocabulary" (p. 59), later identified as Uralex (p. 61). This is an ostensive definition which does not explain what is in the dataset, and which creates a problem for readers unfamiliar with Uralic languages and without access to Uralex. Faced with examples such as *gutna* 'ashes' or *arvi* 'rain' (p. 63), these readers can have no idea of their linguistic status, what their properties are, or of what other words, if any, are related to them. In this particular case, however, the analysis is built on solid linguistic research, so that readers familiar with the philological tradition can take the results on trust, even without a formal explanation.

The key to understanding basic words is found in psycholinguistic speech production research related to the work of Levelt (1989). According to Levelt's model, the speaker selects abstract word forms called lemmas, which are expanded as necessary by morphological rules to form the words found in texts. Corpus linguists work the other way round, and identify sets of related words in texts, these sets also being called lemmas. The shared insight is that the speaker or writer uses basic words stored in the mental lexicon to form the groups of related words found in texts. The term *lemma* is used here as in corpus linguistics. The importance of lemmas in the

study of loanwords is that it is necessary to understand the nature of lemmas both in the donor language and in the recipient language, and also the relationship between the morphological systems.

The expressions *basic word(s)* and *basic vocabulary* correspond to at least two quite different concepts, which have in common the property that they are words from which other words are derived by morphological rules. The first concept concerns words stripped of affixation (and in their written form without punctuation or other markup), which are here called simplex forms, in contrast to complex forms with affixation. The second concept concerns word forms which are used by convention, typically in language teaching, and complemented by strategies for forming related words and recognising them in texts. For example, *hacer* 'do' gives Latin Spanish learners who know the rules access to all the forms of the conjugation, including regular *hacemos* 'we do', and even irregular forms such as *hecho* 'done'. In this way, the inclusion of words in vocabulary lists gives informed language learners access to lists of words.

The relationship between simplex forms and conventional basic words varies from one language to another. English and Malay are remarkably similar in that many content words have simplex forms which are used to form complex words. However, English also has word groups such as *psychiatry*, *psychiatrist* and *psychiatric*, which have no obvious simplex form. The same is true of the Malay borrowings *psikiatri* and *psikiatrik*, which do not fit into the general system. Simplex forms are known in Malay as *bentuk akar* 'root forms' (*bentuk* 'form'; *akar* 'root'), which raises another problem, since *root* is also multiply ambiguous. Semitic roots, for example, are quite unlike Malay bentuk akar, and typically consist of three consonants, e.g. Arabic KTB 'write', and the learner needs to know the rules to generate or recognise morphologically related forms.

A typical lemma is a set of related words which includes the simplex form, e.g. English WALK includes *walk*, *walks*, *walked* and *walking*. The lemma is usually named according to the simplex form, but despite the formal similarity between WALK and *walk*, they are of course logically different objects: WALK is a set of words, and *walk* is a single word. Membership of the lemma is typically restricted by convention, and the English lemma is usually restricted to words of the same part of speech. This excludes the noun *walker* and also *walk* used as a noun from WALK tagged as a verb, and similarly SING includes *sing*, *sings*, *singing*, *sang* and *sung*, but excludes *song*. The use of traditional parts of speech is not necessarily the best criterion to define lemma membership, especially as corpus linguists typically use sets of over 100 grammatical classes for the grammatical tagging of texts. Malay has a rich derivational morphology that frequently changes the grammatical class, e.g. *makan* 'eat' is a verb and *makanan* 'food' is a noun; and lemmas consequently cannot usefully be based on parts of speech. Malay lemmas (whether native or borrowed) here include words derived directly or indirectly from the same simplex form. Dictionaries, incidentally, are likely to tag headwords according to the class of the simplex form, e.g. MAKAN is actually classed as a verb (Kelana & Lai, 1998); but this practice confuses the part of speech of the simplex form with the category to which the lemma belongs.

The distinction between lemmas and their members is important, because corpus word frequency lists can be based on lemmas or distinct lexical items. Different choices result in quite different lists and rank orders (Knowles & Zuraidah Mohd Don, 2004). In the case of published word frequency lists, it is important to know what counts as a word. For example, The Corpus of Contemporary American English (COCA)[1] counts lemmas, while The Academic Word List (AWL)[2] (Coxhead, 2002) "contains 570 word families which frequently appear in academic

[1] https://www.english-corpora.org/coca/
[2] https://www.eapfoundation.com/vocab/academic/awllists/

eISSN: 2550-2131
ISSN: 1675-8021

texts". So-called word families in this context include not only members of the lemma but also other related words, e.g. the headword METHOD includes *methodology* in addition to *methods*.

NATIVE WORDS AND LOANWORDS

Questions concerning what constitutes a loanword arise in an investigation of "9,452 unadapted English loanwords" in a "Corpus of Croatian news portals" (Bogunović, 2023). Unadapted words include simple words such as *leap*, *dusk* and *sweat*, and also the complex *dislike*, *outstanding* and *incredibly*. Adapted foreign words excluded from the investigation include *snowboardi* 'snowboards', *čips*, *tenis*, and *klub* (p. 439). Adaptation seems to involve either modifying the spelling of words to match Croatian conventions, or creating new words by the use of Croatian morphology. A question for unadapted words is what language they belong to. If *Weltanschauung* occurs in an English text, it surely remains a German word, albeit in a foreign context. At the time of writing, *oblast* is being used in news reports, and surely remains a Russian word, even though it is transliterated from область for the benefit of readers unfamiliar with the Cyrillic alphabet. There must be some unexplained criterion to justify the classification of words like *leap* and *dusk* as Croatian words. Frequency of occurrence is presumably involved, and Bogunović excludes hapax legomena, and thus nonce borrowings. It is also important to know how the words are used in the new Croatian context. Bogunović does not explain how words are used once they enter the recipient language, for example whether *dislike*, *outstanding* and *incredibly* are used respectively as a verb, an adjective and an adverb.

Similar problems have been encountered in the study of English loanwords in Malay. Although the concept of loan word might appear to be self-evident, the manner in which English words are used in Malay texts requires an explicit definition of loanwords. In the absence of a totally satisfactory criterion to distinguish English loan words from native English words, the criterion used here is whether or not words appear ('unadapted') in their normal English form. Those that do are English words. Those that do not, having been subjected to Malay spelling rules or morphology, are considered loanwords in Malay. For example, *hotel* in a Malay text remains an English word, whereas *hotel-hotel* or for example *justifikasi* are loan words. This creates some unavoidable anomalies, *import* remaining English and *eksport* becoming a Malay word. If this is unsatisfactory, the problem lies not in the methodology but in the way foreign words are used in real life.

WORD FORMS AND MEANINGS

It has been a commonplace observation since the time of de Saussure (1916), that words are signs that make arbitrary connections – or in more contemporary terms, form the interface – between linguistic form and meaning. A sufficient theory of linguistic borrowing needs therefore to complement the study of the form of loanwords with the study of their meanings. Haspelmath and Tadmor (2009) treat items in word lists as meanings, and use a list of 1,460 meanings (pp. 22-34) to trace loanwords across languages. This view would imply that Swadesh word lists, for example, actually contain meanings for cross-language comparison *mutatis mutandis*, and that many (but not all) word lists are actually lists of meanings.

Although Haspelmath and Tadmor (2009) are undoubtedly on the right track, the emphasis on meaning creates problems for the methodology. Problems of form can be handled by a relational database system, using structured data in tables with fixed relationships. For example, words can

be annotated with a grammatical tag, affixation type and lemma, and *telefon* in a Malay lexical list can be formally associated with its source in an English lexical list. The advantage of this approach is that instead of concentrating on a possibly arbitrary list of loanwords, we can make a systematic study of loanwords in a large dataset. However, the systematic study of meaning cannot be handled by relational logic. Although some semantic relationships can certainly be handled using related tables, and although individual cases can be dealt with *ad hoc*, dealing systematically with meaning would require a huge collection of tables that could not easily be queried using a search engine designed for relational data. What is required is a graph database, which belongs to a quite different type of database management system. This is beyond the capacity of this study in its present form, and so despite the importance of the meanings of loanwords, these things are not dealt with in this paper.

This paper sets out to trace the impact of English loanwords on Malay. They are examined in a corpus of Malay academic magazines in order to assess their distribution in the Malay lexicon and in Malay texts. The general questions guiding the research include:

(1) How frequent are English loanwords, and how is their frequency to be measured?
(2) What form do English words take when borrowed into Malay?
(3) What patterns can be identified in the data to support generalisations about English loanwords in Malay?

DATA AND METHODS

The data for this project is included partly in a corpus and partly in a lexicon. As described above, native English words appear in the Malay text in their English form, whereas true loanwords are modified by Malay spelling or morphology. This clear distinction leaves no orderline cases.

THE CORPUS

The corpus is a collection of Malay texts described by the original compiler as academic magazine articles, and amounting to about two million words. This is roughly half of a bigger magazine corpus, the other half containing Malay non-academic articles. This paper is restricted to the academic articles on the grounds that the others have been insufficiently processed. The focus is on the word list and the patterns extracted from the corpus.

References to academic and non-academic magazines are informal descriptions, and no claim is made that these texts represent academic Malay as a genre (cf Hyland, 2015). To begin with, they are not representative of anything except themselves, and there is a strong bias towards twentieth century linguistics. The academic magazines would appear to be informal in-house publications. The value of this corpus is that it is written by professional writers who produce well-formed Malay written texts containing copious examples of not only Malay morphology but also syntactic patterns to be used for the further development of the MaLex syntactic component.

eISSN: 2550-2131
ISSN: 1675-8021

THE LEXICON

The lexicon consists of a collection of tables organised as a relational database and processed using the language SQL. The loanwords referred to here have been collected in the course of several corpus projects and stored in a digital lexicon. Whereas a conventional dictionary lists distinct words under headwords, the digital lexicon uses related tables, including a table of lemmas roughly corresponding to dictionary headwords, and a table of distinct lexical words that might be listed under the headword in a conventional dictionary. The annotation of the lexical words includes the lemma, which is used as the foreign key to link with the data in the lemmas table. For example, linked to the lemma LOKASI there are eleven related words including *lokasi* 'location', *lokasi-lokasi* 'locations', *dilokasi* 'be located' and *penglokasian* 'process of locating'. Annotations in the lemmas table include the source language, which makes it possible to identify lemmas of English origin (and consequently related lexical words). The advantage of this design is that it goes beyond the inclusion of *lokasi* in a list of loanwords, and traces the penetration of loanwords into the morphology. It also enables separate counts of lexical words and lemmas.

At the time of writing, the Malay lexicon contains altogether 12,496 lemmas, and 49,224 lexical words, or on average about 4 lexical words corresponding to each lemma.

PROCEDURES

Relational logic (Codd, 1970) is independent of the nature of the data. Although it is often used for commercial data, it is also well suited to the handling of linguistic form from speech waveform annotations to meaning representation. The database management system provides the infrastructure for digital linguistic analysis, which in the absence of suitable existing digital models of Malay has to be constructed from the beginning. Processing in the present case begins with identifying and analysing words encountered for the first time in corpus data. Affixes (including clitics) are progressively stripped off until the new form can be assigned to an existing lemma, or else a new simplex form and therefore a new lemma is identified. In each case, words are assigned a grammatical class, and given a rough English translation. Algorithms in this approach do not make independent decisions, but operate in a deterministic fashion to produce output that needs to be checked. This is partly because deterministic rules can produce false positives; for example, if the word *ikan* 'fish' were to be newly encountered, it would be absurdly treated as though the verbal suffix *-kan* were attached to the stem *i-*. The great advantage of this approach is that it generates related datasets containing declarative knowledge of Malay, and thus provides large amounts of information of a kind otherwise unavailable.

Relevant information also speeds up the analysis of new corpus material. In the case of the corpus used here, nearly 98% of the words were already in the lexicon. The remaining 2% corresponds to thousands of new words, the analysis of which is in itself a major undertaking, but which is at least feasible given automatic and automated processing. Loanwords, incidentally, are not given any special consideration in this analysis, although their characteristics may be identified *post hoc*. They have long been recognised diachronically through the failure of expected sound changes, and the synchronic equivalent is exemption from morphological rules. For example, the verb *menelefon* 'telephone' shows the expected modification of the initial consonant of *telefon*, whereas *menteknologikan* 'technologise-' retains it; and in *mentransformasikan* 'transform' the modification is hindered by the initial consonant cluster. As these examples show, some of the phenomena involved in the study of loanwords in Malay are probabilistic in nature.

eISSN: 2550-2131
ISSN: 1675-8021

FINDINGS

The first two subsections here deal with English loanwords in the lexicon, and in the corpus texts. The third subsection deals with English words in the corpus texts which remain English words. Findings for the lexicon are presented in the form of two figures separated by a semi-colon and bounded by round brackets, the first representing the number of occurrences and the second the percentage. Findings for the corpus are presented in the context of the full corpus word list. Again there are two figures, the first representing the rank order and the second the number of occurrences in the corpus. This second figure can be halved to give a rough estimate of the rate of occurrence per million words of text.

ENGLISH LOANWORDS IN THE MALAY LEXICON

At the time of writing, the Malay lemmas table contains 12,504 entries and the Malay lexical words table contains 49,933. On average, the ratio of lexical items to lemmas is about 4:1. Of the lemmas, 3,982 (32%) are borrowed from English, as are 6,794 (14%) of the lexical words, and the ratio of lexical items to lemmas is in this case less than 2:1. The first observation to make is that a large proportion of lemmas is borrowed from English, but that they are much less productive than Malay lemmas as a whole.

It appears that nouns are more prone to borrowing than other word classes (Matras, 2009; Dunn & Vesakoski, 2023), and as expected, most of the loans (3210; 81%) are nouns (van Hout & Muysken, 1994). Many of these nouns (392) end in *-asi*, and correspond to English originals ending *-ation*, e.g. *frustrasi*. A further 30, e.g. *oposisi*, end in *-isi*, and 19 e.g. *infusi*, in *-usi*. In this way, 441 (14%) of the borrowed 3210 nouns correspond to English *-tion* alone.

English verbs are rarely borrowed, only 25 cases having been identified so far. However, denominal verbs can be formed using the circumfix *meng..kan*, e.g. *mengaplikasikan* 'apply'. This circumfix is highly productive, 2042 examples having been recorded so far in the lexicon, and mainly derived from nouns, verbs and adjectives. Of these, 162 are derived from English loanwords, and in this case they are overwhelmingly derived from nouns. (An exact count is not possible, because the class of some loanwords is indeterminate; for example, *meradikalkan* derives from *radikal*, but the English source *radical* can be a noun or an adjective.) Although verbs may not be borrowed directly, Malay morphology seems to be creating equivalent verbs by deriving them from borrowed nouns, including *-tion* nouns.

Most of the loanwords other than nouns (774; 19%) are adjectives, and an interesting finding is that the nature of borrowed adjectives contrasts with that of native Malay adjectives. Typical Malay *kata sifat* (roughly 'adjectives') match the familiar prototype as describing words, and are concerned with permanent or at least long-term characteristics, e.g. *pendek* 'short', *kaya* 'rich', *tua* 'old'. Malay also has a class of words, the full investigation of which remains a task for future research, and which are associated with situations arising from a process or event, e.g. *rosak* 'out of order, kaput', *rebus* 'boiled'. These are related to English past participles, and like past participles, they straddle the fuzzy boundary between adjectives and verbs.

English adjectives are borrowed with their endings, e.g. *fizikal*, *gastrik*, and as these examples show, it can be difficult to explain the meaning without reference to a related noun. In some cases, they are borrowed alongside the corresponding noun, e.g. *deskriptif* alongside *deskripsi*, and *destruktif* alongside *destruksi*. However, whereas the source words belong to the same word family in English ("word family" as used by Coxhead, 2002), the corresponding loan

eISSN: 2550-2131
ISSN: 1675-8021

words in Malay are unrelated, and simply belong to different lemmas. Nevertheless, a case could be made that they belong to an emerging type of word family on the English model. Although English is by far the main source of denominal adjectives in Malay, it is not uniquely so, for example *duniawi* 'secular, of the world' being borrowed from Arabic.

The borrowing of English denominal adjectives complete with their endings creates an apparently new category of adjectives in Malay. Dixon (2010) discusses adjectives in several places, and treats them generally as a small category from a typological point of view. Native Malay adjectives as "describing words" would seem to match the typological prototype for adjectives, and words related to past participles depart from this prototype. In this context, borrowed denominal adjectives depart even further from the prototype, and seem to be something of an anomaly.

This problem has to be addressed in the context of noun modification. Adjectives are the obvious modifiers, but since they are by no means the only ones, it is necessary not to confuse adjectives (a grammatical class or "part of speech") with noun modifiers (which are constituents of a syntactic construction). Malay morphology does not include a form of affixation to derive adjectives from nouns, and where English uses a denominal adjective and a noun, Malay typically uses a head noun and a modifying noun, e.g. *masalah ekonomi* 'economic problem(s)' (*masalah* 'problem'). A reasonable inference is that English adjectives, as in *pemanasan global* 'global warming', are borrowed to compensate for the absence of Malay denominal forms, and in this way borrowing from English is creating a new class of adjectives in Malay.

It may be objected that Malay does indeed have denominal adjectives, including *harian* 'daily', *mingguan*, 'weekly', *tempatan* 'local', and *bandaran* 'pertaining to a town', which are derived respectively from the nouns *hari* 'day', *minggu*, 'week', *tempat* 'place' and *bandar* 'town'. The problem with this claim is that the suffix *-an* used in these cases forms nouns, including nouns derived from other nouns, e.g. *lautan* 'ocean' derived from *laut* 'sea'. The class of the English translation equivalents may cause some confusion, but it is strictly irrelevant in the classification of the Malay words. In the phrase *kerja harian* 'daily work', the head noun *kerja* 'work' is modified by the derived noun *harian*. Nevertheless, words like *harian* seem to straddle the fuzzy boundary between nouns and adjectives. For example, unlike most nouns including *lautan*, and unlike English adjectives such as *daily* and *annual*, they cannot be used to head a noun phrase.

ENGLISH LOANWORDS IN THE CORPUS TEXTS

The corpus word list itself is generally unremarkable and has the characteristics of a typical corpus word list. The first of the figures in brackets below is the rank order, and the second is the frequency. At the top end, the frequency difference between successive ranks begins large but progressively decreases. The most frequent words in the corpus are *yang* 'which, who' (1; 221,470), *dan* 'and' (2; 166,669) and *di* 'at, in' (3; 74,611), which illustrates the expected decreasing frequency differences. At the bottom end, many items tie at the same rank: there are 3,722 items that occur three times, 8,148 items that occur twice, and 33,331 hapax legomena that occur only once.

These most frequent words are, as expected, function words, and the list has to be filtered using a stop list to reveal the content words. The top ten content words are (Table 1):

TABLE 1. The list of content words

Rank	Word	Gloss	Frequency
11	bahasa	language	42108
20	Melayu	Malay	23647
22	orang	person	20746
28	negara	country	18171
38	tahun	year	13946
40	kata	word	12492
41	secara	manner	11363
43	menjadi	become	10989
44	mempunyai	possess	10982
46	Malaysia	Malaysia	10662

These content words are unremarkable ordinary words that can be expected in any written text, and the same is generally true of all the Malay words in the top 200. English loanwords are also content words, and they are generally much less frequent, the top 10 being (Table 2):

TABLE 2. The list of English loanwords

Rank	Word	Gloss	Frequency
76	ekonomi	economy	6695
79	sistem	system	6641
122	teknologi	technology	4421
126	proses	process	4303
140	sains	science	4111
169	politik	politic(al)	3448
176	konsep	concept	3398
179	sekolah	school	3297
195	teori	theory	3041
199	universiti	university	2975

Unlike the Malay words, the English loanwords are technical words that have a natural place in academic writing.

As the top 10 examples illustrate, English loanwords are most commonly used in the simplex form. Less than 1350 distinct complex forms have been recorded so far, and over 800 of these (60%) are hapax legomena. The top ten are (Table 3):

TABLE 3. English loanwords commonly used in the simplex form

Rank	Word	Gloss	Frequency
408	kritikan	criticism	1556
803	perindustrian	industry	699
1405=	pengkritik	critic	364
1676	aspek-aspek	aspects	292
1759=	perakaunan	accountancy	273
1841=	berkualiti	of quality	259
1860=	menganalisis	analyse	256
1910=	faktor-faktor	factors	248
1947	berkomunikasi	communicate	243
2079	mengkritik	criticise	220

Except in the case of the first two of these, several items share the same rank. These figures indicate that although the morphological infrastructure is in place to enable the formation of complex forms, the potential is not much exploited in this corpus. No explanation, incidentally, can be offered at this stage for the frequency of the lemma KRITIK in this list.

ENGLISH WORDS IN THE CORPUS TEXTS

This subsection deals with English words that occur unchanged in the Malay texts. They remain unchanged because the English spelling is also well formed in the Malay spelling system. There are 5,394 items listed, and of these 2,886 (54%) are hapax legomena. Nearly all are simplex forms, but there are some rare exceptions, including *windows* (4074; 79) and *states* (8748; 21), which have plural marking.

Although dealing with these English words might seem straight forward in principle, this is not always the case in practice. This is because a number of Malay words have homographs in English, the most frequent cases being *lain* 'other' (55; 9597), *pun* 'also' (99; 5441), and *air* 'water' (137; 4045). The less frequent *jam* 'clock, watch' (817; 681) and *fail* 'file' (896; 626) prove to be Malay words in the corpus context. The spelling "air" is potentially ambiguous and is used as an English word in the context of Malaysian petrol stations which provide air for tyres. The Malay spelling "fail" is potentially confusing as it corresponds to English "file". Some other items are English function words: *the* (242; 2598), *of* (325; 1953) and *a* (346; 1830). These are presumably not borrowed on their own account but are included in English expressions used in texts.

Leaving aside irrelevant homographs and function words, the top 10 English words are (Table 4):

TABLE 4. The list of irrelevant homographs and function words

Rank	Word	Gloss	Frequency
305	novel	novel	2081
397	media	media	1601
485	data	data	1311
489	program	program	1296
645	idea	idea	935
680	era	era	877
699	bank	bank	850
762	model	model	755
764	bin	n/a	753
795	moral	moral	713

With the obvious exception of *bin*, these are words one might expect to be used by academic writers. *Bin* 'son of' is used in Arabic names (most infamously in *Osama bin Laden*) and is another irrelevant homophone that found its way into the English list.

DISCUSSION

This discussion section first deals with two issues that have already arisen, namely words and word sets, and the impact of loanwords on the recipient language structure. The third subsection discusses the social acceptability of loanwords as a new issue arising from the previous sections.

eISSN: 2550-2131
ISSN: 1675-8021

WORDS AND WORD SETS

Distinguishing lemmas and lexical words has long been standard practice in corpus linguistics, and in this respect loanwords are treated like any other words. A separate linked lemmas table proved extremely useful in finding patterns in the data. More generally, it is important to organise linguistic information in such a way that useful searches can be made and insightful results retrieved from the data. This approach is more systematic than a list of words of ill-defined status, and perhaps it deserves to be the normal approach in the study of loanwords.

The corpus word frequency list proved to be less enlightening than the lexicon, but at least it confirmed the expectation that English loanwords tend to be technical in nature. It also indicates that the potential for complex loanwords is in practice little exploited. The word list could have been reorganised as a list of lemmas, but a preliminary investigation indicated that this would not in practice yield much new useful information.

Coxhead's word families (2002) deserve more attention than it has been possible to give them here. They are presumably sets of lemmas, and contain words that are formally related, so that *nation* and *national* belong to the same family, whereas *enemy* and *hostile* do not. English word families are related by morphology, but Malay does not (yet) have the morphological infrastructure that connects *psychiatry* to *psychiatrist* and *psychiatric*, so that when borrowed into Malay these words have to be regarded as unrelated. It is possible that a single corpus of two million words is just far too small a data set to detect possible patterns in this case, and word families will have to be reconsidered with a greatly enlarged data set.

IMPACT ON THE STRUCTURE OF MALAY

Loanwords are conventionally regarded as items transferred from one language to another. This may be true of Malay loanwords in English, including *amuk* 'amok' and *gedung* 'godown, warehouse'. It may also be true of English loanwords in Malay that are used in everyday popular speech, including *kopi* 'coffee', *kek* 'cake' and *karipap* 'curry puff'. A different explanation is required for the findings reported above relating to texts written by academics, and for that matter also by journalists and other professional writers. The transfer takes place not between language systems but in the brains of bilinguals. Malay speakers who also know English can draw on the lexical resources of both languages, and it is not surprising if an English form is used or adapted when it is more convenient than a native Malay one, if indeed a Malay equivalent exists.

The process of borrowing involves discontinuous change, which involves a change of state for which there is no precise time when the change takes place. For example, English *shire* was largely replaced by the French form *county* in the medieval period, but there is no discernible time when *county* became an English word. Personal opinions are likely to differ concerning which of the loanwords discussed above have truly become Malay words, and which are still best regarded as English. Although discontinuous change cannot be assigned a time, there may be observable indications that it is taking place, or has already done so. Native spellings and morphology are weak indications, and may be disregarded. (For example, the occurrence of the phrase *four oblasts* in an English text does not mean we have to add *oblast* to the list of Russian loanwords in English). Much more important are emerging patterns in the data.

Malay has a subclass of nouns ending *-si* derived from English nouns ending *-tion*. This has not only greatly increased the stock of abstract nouns in Malay, but has also added to the class of nouns which can form derived verbs with the circumfix *meng..kan*. The structural developments

eISSN: 2550-2131
ISSN: 1675-8021

brought about by borrowing *-si* words are thus unrelated to the properties of *-tion* words in English. Whereas English gives logical priority to verbs, so that the abstract noun *location* is derived from the verb *locate*, it is the other way round in Malay, so that the verb *melokasikan* is derived from the abstract noun *lokasi*. The borrowing of complex adjectives has created a new possible subclass of adjectives in Malay, which also extends the range of noun + adjective constructions as an alternative to noun + noun constructions, and thereby simplifies the translation of English adjective + noun expressions into Malay. Many borrowed English adjectives are singletons, unless other members of the same English word family are also borrowed. However, if the use of denominal adjectives becomes more widespread, and if new denominal adjectives can only be obtained by borrowing from English, alternative sources could eventually be developed. The interaction between English loanwords and traditional Malay morphology has led to what could prove to be fundamental changes in the three major grammatical classes, nouns, verbs and adjectives.

THE SOCIAL ACCEPTABILITY OF ENGLISH LOANWORDS

This paper has sought to make an objective assessment of the impact of English loanwords on contemporary Malay. However, it is also necessary to consider the social and political impact of English loanwords. Opinion has for centuries been divided between those who believe that loan words are necessary, and those who seek to deprecate them. In the case of English, Sir Thomas Elyot (1531) borrowed a number of Latin words (including *education*) into English "for the insufficiency of our own language", while the opposing view was expressed in 1557 by Sir John Cheke, professor of Greek at Cambridge, who argued that "our own tongue should be written clean and pure, unmixed and unmangled with borrowing of other tongues" (Knowles, 1997, p. 70). Elyot's position reflects a situation in which increasing literacy was creating a class of people who were literate, but who could not read texts in the original language. Cheke, by contrast, reflected the view of the élite who could read Latin, and in his case also Greek.

Elyot and Cheke have had many followers, and disagreements over loanwords have been repeated many times in different cultures, including present-day Malaysia. Although some people may not approve of English loanwords in Malay, the relationship between the English and Malay languages is not the central issue. The central issue is the relationship between Malay culture and global culture. If Malay is to be regarded as a significant global language, then it requires the means to discuss matters of international concern, ranging from the environment to international relations and scientific and technological innovations. This cannot realistically be done using the kind of Malay that has been inherited from past centuries.

The parallelism of borrowing from Latin into English and from English into Malay raises two caveats. The first is that the borrowing of Latin words by literate English people in the sixteenth century obviously had nothing to do with the Roman Empire. Likewise, the contemporary borrowing of English words by literate Malay speakers has nothing to do with the defunct British Empire. Latin and English were respectively the conduits which enabled the transfer between cultures. The second and related caveat is that linguistic influence in either case does not involve the wielding of power over the unwilling. Latin writers whose work has remained influential over the centuries include Jerome from Dalmatia in modern Croatia, and Augustine of Hippo, who was of Berber origin. Both writers lived in places formerly under Roman rule, and both chose to write in Latin. The anonymous people who now introduce English words into their first language are likely to be bilinguals who seek to strengthen their own language, and this includes those who borrow English words into Malay. The proliferation of English words in Malay

eISSN: 2550-2131
ISSN: 1675-8021

and other languages has everything to do with the globalised world order created in 1945, and the need to enable local languages to take their place in the globalised world.

CONCLUSION

The study of loanwords has been rejuvenated in recent decades by the adoption of techniques developed in corpus linguistics, which have made it routinely possible to study large data sets containing loanwords in context. The extension into relational databases enables the computation of linguistically organised findings which constitute declarative information about Malay. Importantly, this declarative information can be understood by anyone who has either linguistic expertise or a knowledge of Malay. Although this paper concentrates on loanwords, this is just one demonstration of what can be done with corpora and relational databases, leaving aside the potential of future developments.

The generation of declarative linguistic information is important in view of the ever-increasing demand for computer-readable information of all kinds for teaching and research purposes. For example, countries across the world face the task of upgrading their language education systems by aligning them with the *de facto* international standard, which is the Common European Framework for Languages (CEFR). Malaysia has already done this for English, but the same is required for Malay and other Malaysian languages used in education. This requires knowing which words are appropriate at each stage for L1 and L2 learners as they progress from beginning earners to intermediate and advanced learners. As this paper has demonstrated, education is a field in which loanwords take root and thrive, and these word lists will accordingly have to include English loanwords. Apart from words introduced for the development of general proficiency, learners need to acquire words for special purposes, particularly in the teaching and learning of maths, science and technology, and many and perhaps most of these words will be loanwords from English. Since most Malaysians, including teachers, do not know Latin or Ancient Greek, which are the traditional sources of technical and scientific words, the language of English loanwords is itself a topic to be addressed, irrespective of the language of instruction.

The borrowing of English words in contemporary Malay goes far beyond the traditional borrowing of words through language contact. Words and expressions are coined for English in the first instance, and from English they spread to languages across the globe. In the globalised world, it would be inappropriate for each language to invent its own terms for innovations and new concepts, because that would hinder communication where it is most necessary. By contrast, borrowed forms adapted to the conditions of different languages enable important concepts, words and expressions to be understood across languages. For example, the borrowed *semikonduktor* is in practice more useful to the Malaysian semiconductor industry than some invented form based on native Malay words. Malaysia is plugged in to the globalised world, and its culture and its national language need to be responsive to developments in the outside world. This is achieved, at least in part, by borrowing words from English. English loanwords play an essential role in maintaining and enhancing the relevance and position of Malay as a significant language in the globalised world.

eISSN: 2550-2131
ISSN: 1675-8021

REFERENCES

Biber, D., & Reppen, R. (2015). *The Cambridge Handbook of English corpus linguistics* (Eds.). Cambridge University Press.

Bogunović, I. (2023). A corpus-based approach to English loanwords: Introducing the database of English loanwords in Croatian. *Fluminensia*, *35*(2), 437-460.

Bond, C. (2025). Corpus linguistics as a research method in nursing: A practical approach to analysing language data. *JAN*, *81*(8), 6960-6967.

Codd, E. F. (1970). A Relational Model of Data for large shared data banks. *Communications of the ACM*, *13*(6), 377-387.

Coxhead, A. (2002). The academic word list: A corpus-based word list for academic purposes. *Language and Computers*, 73-89.

de Heer, M., Blokland, R., Dunn, M., & Vesakoski, O. (2023). Loanwords in Basic vocabulary as an indicator of borrowing profiles. *Journal of Language Contact*, 16, 54-103.

de Saussure, F. (1961). *Cours de Linguistique Générale.* Payot.

Durkin, P. (2014). *Borrowed Words: A history of loanwords in English*. Oxford University Press.

Fries, C. C., & Traver, A. A. (1960). *English word lists*. George Wahr Publishing Co.

Haspelmath, M., & Tadmor, U. (2009). *Loanwords in the World's Languages: A comparative handbook* (Eds.). Mouton de Gruyter.

Havumetsa, N.(2023). Lexical borrowing in journalism in a time of political crisis. *Perspectives*, *31*(3), 562-575.

Hyland, K. (2015). Corpora and written academic English. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics* (pp. 292–308). Cambridge University Press.

Kelana, C. M., & Lai, C. (1998). *Kamus Perwira*. Penerbitan Daya Sdn Bhd.

Knowles, G. (1997) *A Cultural History of the English Language*, Arnold.

Knowles, G & Zuraidah Mohd Don (2004). The notion of a lemma: Headwords, roots and lexical sets. *International Journal of Corpus Linguistics*. *9*(1), 69-81.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.

Matras, Y. (2009). *Language Contact*. Cambridge-New York: Cambridge University Press.

Michaud, M. & Hollenback, M. (2015). Material foreign loanwords and the emergence of English Japanese. *Kwansei Gakuin University Humanities Review*, 20, 259-273.

Morrow, P. R. (2020). Japanese loanwords in English: A corpus-based study. *Journal of Nagoya Gakuin University*, *57*(1), 1-13.

Mulcaster, R. (1582). First part of the elementary vvhich entreateth chefelie of the right writing of our English tung, set furth by Richard Mulcaster. In the digital collection *Early English Books Online*. University of Michigan Library Digital Collections.

Portugal, E., & Nonnenmacher, S. (2024). Every world is a world: loanword ideologies and linguistic purism in post-Soviet Armania. *Multilingua*, *43*(3), 331-364.

Oh, Y., & Son, H. (2023). Lexical borrowing in Korean: A diachronic approach based on a corpus analysis. *Corpus Linguistics and Linguistic Theory*, *20*(2), 407-431.

Richard, P. (2024). The global spread of English loanwords: Implications for linguistic diversity (May 10, 2024). Available at SSRN: https://ssrn.com/abstract=4823941

Tadmor, U. (2009). Loanwords in Indonesian. In Haspelmath, M., & Tadmor, U. (Eds.), *Loanwords in the World's Languages: A comparative handbook* (pp. 686-716). Mouton de Gruyter.

Thomason, S. (2021). *Language contact*. Edinburgh University Press.

Trench, R. C. (1851). *On the study of words*. Macmillan.

van Hout, R., & Musyken, P. (1994). Modeling lexical borrowability. *Language Variation and Change*, *6*(1), 39-62.

Zenner, E., Rosseel, L., & Calude, A. (2019). The social meaning potential of loanwords: Empirical explorations of lexical borrowing as expression of (social) identity. *Ampersand*, 6. https://doi.org/10.1016/j.amper.2019.100055

Zuraidah Mohd Don & Knowles, G. (2020). New tools for old tasks: A new approach to the investigation of Malay. JALA. *2*(3), 21-38.

ABOUT THE AUTHORS

Zuraidah Mohd Don is a Professor at UCSI University. She has published many articles in peer-reviewed journals in Language and Linguistics, including articles arising from collaboration with Dr Knowles *inter alia* on MALEX and other corpus projects. Her most recent publications are co-authored books on CEFR-aligned assessment at the tertiary level.

Gerry Knowles is a retired computational linguist from the UK, and started the MaLex project in collaboration with Professor Zuraidah, having no previous knowledge of Malay. He has since developed the computational infrastructure which supports this paper, drawing on the linguistic expertise of Professor Zuraidah.

Nor Shahila Mansor is an Associate Professor and currently serves as Head of Department at the Department of Asian and European Languages, Faculty of Languages and Linguistics, Universiti Malaya. Her expertise includes Spanish language and linguistics, Translation and Interpretation, Sociolinguistics, and the Teaching and Learning of Spanish as a foreign language.

eISSN: 2550-2131
ISSN: 1675-8021