WORD FREQUENCIES AND BIGRAMS IN BAHASA MELAYU

Zuraidah Mohd Don
Faculty of Languages and Linguistics
University of Malaya

Gerry Knowles
University of Lancaster

Abstract

This paper reports on some preliminary findings on word frequency in the MALEX database. The most frequent words are described, attention being paid to the position of content words in the frequency list. Nouns emerge as the most problematic
to a particular set of genres, or even to a particular text. Bigrams are studied both as sequences of individual words and as sequences of grammatical tags. Whereas the tag sequences reflect syntactic rules and thus the hierarchical structure of syntax, sequences of individual words reflect quite a different kind of linear structure which has begun to emerge in recent years in corpus linguistics.

Introduction

The aim of the research reported in this paper is to find out what kind of words, expressions and grammatical constructions are normal in the Malay language. Much effort in linguistics is devoted to finding out what is possible within the bounds of a particular theory; but it is also important to know what is normal. Language teachers and language learners, for example, need to know common words and expressions before rare and unusual ones. In developing software for natural language processing, it is important to know what is likely to occur next at any point in a sentence. As far as we know this is the first time frequency information has been made available for Malay,

although Muhadjir *et al.*, (1996) have published lemmatised word frequency tables for Indonesian.

Our findings are based on the preliminary analysis of a tagged corpus of Malay texts. We begin with the frequency of individual words, then of individual word classes, and we have made a start on the study of strings of words or n-grams. A concordance package such as WordSmith can identify strings associated with particular words, e.g. that *baik* is found in expressions such as *yang baik* or *yang tidak baik*, but this is only for one word at a time. What we are doing is to examine the lexicon as a whole and look for recurring strings. Our study so far is limited to bigrams, or sequences of two words. Since bigrams can be considered as strings of individual words, e.g. *tidak baik*, or as strings of grammatical classes, e.g. negative particle + adjective, we have used the expanded terms *word sequence* and *tag sequence* in the main text.

The MALEX database

The Malex database is the outcome of the analysis of a corpus of roughly 2.3 million words containing political speeches, novels and newspaper articles[1], and a few miscellaneous texts. It is essentially a lexicon designed in the form of a relational database. The main table contains over 30000 orthographic forms each with an associated grammatical tag identifying its grammatical class. These tags are based on a study of word class in Malay discussed in detail elsewhere (Knowles and Zuraidah, in press a), for the present purpose we shall treat *kata nama* as roughly equivalent to 'noun', *kata kerja* to 'verb', and *kata sifat* to 'adjective' The tags in the main table are linked to a tagset contained in a related table.

Each lexical entry includes a frequency field to record the number of times it occurs in the corpus. To obtain a word frequency list, the whole corpus is scanned, and a word count recorded in the frequency field for each lexical entry The main table is then ordered according to frequency, so that the words appear in their rank order. The raw list is not essentially different from a list generated by a concordance package such as WordSmith, but the existence of related tables links the list in the database table to a wide range of information. Individual word frequencies can, for example, be pooled to count the frequencies of grammatical tags, so that we can also rank the tags according to their frequency in the corpus.

---

[1] We are pleased to acknowledge the generosity and support of Dewan Bahasa dan Pustaka both in enabling us to develop the original tagset, and in providing us with large amounts of data from their archives.

The main table is linked to the corpus texts through a spelling normalisation table. Words can appear in texts with an initial capital letter, and may occasionally be spelt incorrectly; and it is important to remove these irrelevant details when counting words in the lexicon. This normalisation table is also used when counting bigrams. A (specially written) computer program scans the words of the corpus texts, and for each word looks up its standard orthography in the spelling normalisation table, and then looks up its grammatical tag in the main table.

Word frequencies

There are many areas in which a basic knowledge of word frequency is essential. The language learner, for example, needs to acquire the most frequent words of the language in the early stages. The speech engineer needs to know what words are most likely to be needed for speech recognition, and lexicographers need to know what are the most important words to include in a dictionary. Linguists have not in general used frequency information in linguistic description, although it is becoming increasingly available in corpus linguistics.

By far the most frequent word in the corpus is *yang* (72715) followed by *dan* (63370) and *tidak* (31047). Positions 5, 6, and 7 are occupied by *ini* (28294), *di* (27957) and *itu* (25756). What is noticeable about these figures is the steep drop in frequency in the first three words. Twenty words have a frequency above 10,000, and 255 above 1000, and 2222 above 100, so that the overall slope is roughly linear.

The most frequent words are in general, not surprisingly, function words. There are two content words with a frequency above 10,000, and a total of seven content words in the first fifty words in rank order, at which point they begin to become frequent. The most frequent verb is *menjadi* (6113) in position 39, and the highest ranked adjective is *baik* (3209) in position 59

The frequency of function words, along with most verbs and adjectives, can be assumed to be to some extent independent of the genre or the subject matter of the text. Nouns, by contrast, are closely bound to subject matter. The most frequent noun is *negara* (13304, rank 14) followed by *orang* (11106, rank 15) and *kerajaan* (8099, rank 27), *Islam* (7166, rank 32) *Malaysia* (6987, rank 35) and *rakyat* (4556, rank 49). Apart from *orang*, which has a partly grammatical function as a classifier (e.g. *dua orang lelaki*, *seorang budak kecil* etc.), all these nouns very clearly reflect the political speeches that make up over half of the corpus. The political context also explains the high frequency of *tuan-tuan* (1400, rank 161), and *puan-puan* (1274, rank 187), routinely used as a form of address at the beginning of a speech.

The distorting effect of text type is also found in the frequency of pronouns. By far the most frequent is *kita* (30873) as high as rank 4, used in the political speeches to include *kerajaan* and *rakyat*. The contrasting pronoun *mereka* (18122) is also high at rank 10. (The exclusive *kami*, on the other hand, occurs only 1279 times and comes in at rank 186.) *Saya* (10046), sometimes guessed to be the most frequent word in the language, comes in at rank 18, and is followed by *dia* (9459, rank 23) and *aku* (2402, rank 88). For the second person, *engkau* (1834, rank 125) comes way ahead of *anda* (64, rank 3083); although *anda* is the form recommended to foreign learners of Malay, it is actually less common than the English word *you* (156, rank 1593). These second person figures reflect usage not in the political part of the corpus, but in the novels.

These findings are not in any way surprising, because in compiling our corpus we have made use of texts that happened to be available, without any attempt to make (or indeed any possibility of making) a balanced selection of texts. To obtain undistorted figures, we need a much wider range of text types from which we could take a representative sample of perhaps a million words. Such a sample would be highly desirable, but until it becomes a possibility, we have to make do with what is available. In evaluating the results reported below, we have to take into account the possible distorting effects of a biased sample.

Pooled word frequencies and lemmas

When dealing with word frequencies, it is common practice to pool the counts for individual words under a single headword or lemma. For example, rather than calculate separate frequencies for *walk*, *walks*, *walked*, and *walking*, we might count the occurrences of the verb **walk** as a whole. For English, it is probably more useful to deal with the frequency of the lemma than of the individual members of the lemma.

In view of the very different structure of the Malay lexicon, the lemma in Malay is not organised in the same way as in Indo-European languages (Knowles and Zuraidah, 2004), and contains words related by derivational morphology rather than inflection. We accordingly have to check that pooling the frequencies of individual words is a meaningful exercise. In some cases it is clearly helpful. For example, *ada* is a good example of what might be thought of as a frequent Malay word. As an individual word, it occurs 9345 times, and is ranked 24. However, it is supported by other frequent words, including *adalah* (6864, rank 36), *keadaan* (1990, rank 111) and *berada* (829, rank 328) and some rarer words such as *mengada* and *diadakan-adakan*, both of which occur only once. As a lemma, **ada** occurs 20965 times, and is the ninth

most frequent lemma. Other cases are not so straight forward. For example, *kerajaan* belongs to the lemma **raja**, which also contains among its members *raja*, *raja-raja*, *rajanya* and *diraja* on the one hand, and *kerajaannya*, *kerajaanlah*, *kerajaan-kerajaan*, *berkerajaan*, *pro-kerajaan*, and *anti-kerajaan* on the other. It is not immediately obvious whether the 'king' group and the 'government' group should be treated together or separately (and in the latter case what the theoretical grounds for the sub-grouping should be) The frequency patterns of lemmas need more investigation and will not be discussed further here.

Tag frequencies

From the main MALEX table we know how often each word occurs in the corpus, and we have a grammatical tag for each word. This means it is easy to calculate the frequency of occurrence of individual tags in the corpus. Some 33% of words in running text are nouns, and a further 18% are verbs, so that over half the words of the text are nouns or verbs. 10.8% of words are conjunctions, a figure that seems remarkably high until one remembers it includes the two most frequent words of all, namely *yang* and *dan*, together with some other high frequency words such as *tetapi* (9917, rank 20), *kerana* (8722, rank 25) and *atau* (7101, rank 33). Words we have tagged as kata sendi account for 8.5% of the data, but here there may be an important difference between Malay classes and English ones. Although English co-ordinating conjunctions and prepositions are regarded as quite different parts of speech, it is quite possible that on further analysis kata penghubung subordinat and kata sendi will prove to be members of a single superclass. Some words such as *dalam* and *atas* which we have tagged as kata nama lokatif can also pattern like kata sendi.

Kata sifat 'adjectives' are surprisingly rare at only 6.7%. If we assume that every adjective is in some way attached to a noun (i.e. following the patterns *rumah besar*, *rumah yang besar* or *rumah itu besar*) there are only enough adjectives for one noun in five. In practice we know that some adjectives are used in adverbial expressions of manner, so that in reality even fewer nouns have associated adjectives. This is an important finding, because when linguists invent data containing noun phrases, they very often invent noun phrases containing adjectives: these constructions are actually less common – and to that extent less 'normal' – than linguists might like to think.

Turning now to the subcategories of the major classes, we find that most nouns are ordinary common nouns, kata nama am. One of the special characteristics claimed for Malay is the use of penjodoh bilangan or 'classifiers' in expressions such as *sebuah kereta* or *sebiji telur*. However, only 0.09% of

words in running text are tagged as classifiers at all. Now if 33 in a hundred words of text are nouns, and only one of these is a classifier, then it is quite impossible for more than one noun in 32 to be preceded by a classifier. Since we know that in practice words such as *buah* and *biji*, along with *orang*, are frequently used as head nouns, the number that actually function as classifiers must be very much lower. Further research will be needed to find out how rare this special characteristic of Malay really is.

The 18% of the text devoted to verbs includes figures of 8.6% active transitive verbs, and a further 2.1% passive verbs. The majority of verbs are thus transitive. Intransitive verbs account for only 2.7% of running text, or about one verb in seven. When tagging Malay words, there are many words for which it is difficult to decide whether they should be classed as intransitive verbs or adjectives. Some Malay linguists (such as Abdullah, 1974) do not make any distinction at all. We believe there are good reasons for maintaining the distinction, but the evidence certainly merits further consideration. A subclass of verbs that turned out to be relatively common at 1.4% of text are verbs formed by the addition of the prefix *beR-* to a noun, of which the most frequent is *berjaya* (1911, rank 119). These are an interesting set of verbs, and are sometimes followed by an adjective, e.g. *bernilai tinggi*.

Malay verbs are of course not marked for tense, aspect or mood, and separate words are used for the purpose. Less than one verb in seven is accompanied by a marker for tense or aspect, and about one verb in thirteen is has an associated marker for mood. It is difficult to assess these findings without having comparable figures for finite and non-finite verbs in English, but it does seem clear that marking these categories is very much the exception rather than the norm.

Tag sequences

Tag sequences are used in English corpus linguistics to determine the tag in cases where a word can have more than one possible tag. For example, *telephone* can be a noun or a verb, but it is more likely to be a noun when it follows an article, as in *the telephone*, and more likely to be a verb when it follows a modal verb, as in *must telephone*. Some preliminary work was carried out to see if such sequences could be of some use in tagging Malay texts, but they proved to be of no use whatsoever. On the other hand they do throw some light on Malay syntax, and are therefore of some use in the development of a parser.

In view of this role in parsing, it is important to be careful when dealing with syntactic boundaries. In the study of collocations, it is enough that words are adjacent or at least close to each other. But in the case of syntax, we would

not expect any connection between the tags of the word at the end of one sentence and the word at the beginning of the next, or between the tags of words on either side of a semicolon or even a comma. For that reason, words followed immediately by punctuation were classed together as in 'final position'

To obtain the statistics for tag sequences, the corpus was searched, and the tags for each word followed by the tag for the next word, or by the symbol "*" to mark words in final position. The probability was then calculated of the occurrence of the second tag (or "*") given the first tag. The probabilities are more clearly expressed as percentages. For example, given a tag N there is (in our corpus) a probability of 0.29 that the next tag will also be N, or in other words, 29% of nouns are followed by another noun.

14.5% of nouns are followed by a conjunction (in many cases of course *yang*), and a further 13.5% are in final position. This means that over half of all nouns (57%) occur in just three environments. Only 7.9% of nouns are followed by a verb, and only 7.5% by an adjective. This could be because noun phrases end with words such as *ini* or *itu*, but in fact only 4.9% of nouns are immediately followed by such words. These figures, together with the overall infrequency of adjectives noted above, indicate that adjectives are not the default modifier of nouns as one might expect. It is much more common for a noun to be followed by another noun, and in many cases the noun will be the modifier of the first.

In view of the overall frequency of nouns, and the frequency of transitive verbs, it is not altogether surprising that 36% of all verbs are followed by a noun. The 14.8% of verbs followed by a preposition are unlikely all to be intransitives or passives, and the 11.3% of verbs in final position are also unlikely all to be intransitive. To identify more detail, we have to subcategorise the grammatical tags. In fact only 3.1% of all verbs are intransitives followed by a preposition, and there are over twice as many transitives in this position, with rather more passives than actives. (After a passive verb the preposition will of course in many cases be *oleh*.) In final position there are more transitive verbs (3.3% of all verbs) than intransitive (2.8%). To account for these figures we have to anticipate that a large proportion of the objects of transitive verbs are actually omitted. An interesting statistical detail is that 3.7% of verbs are followed by an adjective: this is in fact the commonest means of modifying verbs, for adjectives immediately following a verb function as manner adverbials.

The extra detail provided by the inclusion of tag subclasses throws some useful light on the distribution of prepositions, and raises some questions. Most prepositions (62%) are grouped together in a general class, while the remaining 38% are locative prepositions such as *di* and *ke*. 42% of the general prepositions and 50% of the locative prepositions are immediately followed by a general noun. A further 17% of the locative prepositions occur immediately

before a locative noun to form phrases such as *di dalam* and *ke atas*. This means that expressions such as *di rumah* are three times as frequent as expressions such as *di dalam rumah*. (What the study of bigrams cannot reveal, of course, is the frequency of expressions such as *dalam rumah*, i.e. without the locative pronoun *di*.) General prepositions are also frequently found immediately before verbs, possible examples including verbal expressions introduced by *dengan* or *untuk*. Nearly all the verbs involved are transitive. 15% of all general prepositions are followed by an active transitive verb, compared with only 1.5% before an intransitive verb, and only 0.08% before a passive verb.

Word sequences

In order to identify the most frequent sequences of individual words[2], the corpus has to be searched and a record kept of each pair consisting of a word and the following word. Given a lexicon of tens of thousands of words, the number of possible pairs runs into hundreds of millions, and many of these pairs will occur only once and will therefore be of limited interest. A more realistic way forward is to set out to identify a fixed number of the most frequent word pairs, and for this research we looked for the most frequent thousand pairs.

The most frequent combinations necessarily include the most frequent words, and so the corpus was searched working down the rank order in the lexicon. First, all pairs beginning with *yang* were put into a table and ranked for frequency. All entries after the first thousand were then deleted. Then pairs with *dan* were then found and merged with the *yang* pairs, ranked for frequency and all except the first thousand deleted. As more words are processed in this way, the pairs table gradually approximates to the true set of the most frequent pairs in the corpus. In the event, the last word to contribute a pair to the table was in rank position 989 in the lexicon table.

From the examination of the pairs table, it is immediately obvious that there are two different factors affecting the formation of frequent pairs. The first is the frequency of the individual words of the pair, and the second is syntax. The most frequent pair is *yang tidak*, containing the words ranked first and third respectively; and it is also the case that the rules forthe formation of relative clauses allow *tidak* to follow *yang*. By contrast, the rules of Malay syntax exclude the possibility of *tidak yang* occurring as a pair at all. The fact that the syntax allows words to follow in linear order does not mean of course

[2] We are pleased to acknowledge the support of Open Horizon in carrying out this research.

that they are grammatical units. *Yang tidak* is not a unit, while *tidak ada* (rank 2) is; *kita tidak* (rank 3) is not a unit, while *negara kita* (rank 4) is. Overall only 22% of the pairs form identifiable grammatical units.

Closer inspection reveals that the grammatical status of the pairs depends on whether the words would be classed as 'function' words or 'content' words. The biggest single group consists of function words, and very few of these form grammatical units. Nearly all the exceptions are of the type *di mana* (rank 40), and include words like *ini* or *itu*, or *sini* or *sana*, as the second member of the pair. At the other extreme, 56% of pairs containing two content words, e.g. *orang Melayu* (rank 32), form grammatical units. Although only 23% of cases in which a function word is followed by a content word do so, these tend to fall into two recognisable categories. One of these consists of *yang* and an adjective, e.g. *yang lain* (rank 37), and the other of a locative preposition and a locative noun, e.g. *di dalam* (rank 7). Less than 13% of pairs containing a content word and a function word form grammatical units, and most of the ones that do have *ini* or *itu*, or a personal pronoun, as the second word.

Discussion

The initial findings of these frequency studies draw attention to the linear structure of language. Behind the figures are probabilities that having selected a particular word or type of word, the speaker or writer will go on to select some other particular word or type of word. This linear structure co-exists with the hierarchical structure that has always been the main focus of syntactic study.

If we take a top-down approach to syntax, starting with the sentences and drawing tree structures for them, it is tempting to give logical priority to the hierarchical structure, and see the linear structure as the outcome of mapping the hierarchical structure on to a linear string of words. The problem is that it is difficult to see how this procedure could relate to any normal use of language. One cannot, for example, imagine a speaker or writer constructing a tree and then selecting suitable words to match the leaves. It is easy enough to write a computer program to generate arbitrary grammatical trees and select lexical items, but this has nothing to do with communication, and thus rather misses the fundamental point of human language.

If, on the other hand, we take a bottom-up approach, we start with words and word strings, and trace the way they fit together to form phrases and sentences. The hierarchical structure is something we infer from recurring distributional patterns in the data. In developing a parser for the corpus, for example, we have to identify recurring patterns, and then ascertain the nature

of the higher level category which they instantiate. If we parse sentences in conventional written texts, the parser will trace the structure all the way up to the level of sentence; but that is because such texts are by convention composed in complete well-formed grammatical sentences in the first place. If we parse conversational texts, we can expect well formed phrases, but the sentence structure may well be incomplete. The bottom-up approach is the one consistent with an empirical, data-driven analysis of natural language texts used in real communication.

Linear strings of words are what we actually encounter in texts, whereas hierarchical structures are theoretical constructs. The important point that emerges from the study of frequencies is that the kind of grammatical structure linguists are trained to think about form only a subset of the recurring patterns to be found in linear strings of words. This is why we have used the vague term *grammatical unit* above. The kind of word pairs that stand out in the table are expressions such as *di dalam* or *di Malaysia*, which are necessarily grammatical constituents. An expression such as *orang Melayu* looks like a constituent and is likely to function as one in many cases; but when words are added to it, e.g. to form *ramai orang Melayu* or *orang Melayu itu*, more than one bracketing is possible, and it is possible to draw more than one kind of tree. Native speakers of Malay are likely to want to group *orang* with *Melayu* as a single unit. On the other hand, in a phrase such as *tiga buah kereta*, the classifier *buah* is more likely to be grouped with the numeral *tiga* than with the noun *kereta*. There is no fixed or objective constituency structure in such phrases, and much depends on the tagging (e.g. whether or not *orang* is treated as a classifier), and on the design of the parser, e.g. whether it gives the whole phrase a flat parse or looks for sub-groupings within it.

A quite different principle which governs the co-occurrence of words in texts is collocation, the association of words according to their meaning. In collocation studies it is normal to look for collocates several words to the left and to the right of the key word. In examining only sequences of two words, we are in effect looking only for collocates which are immediately adjacent to the key word. Now if we examine our word sequences from a conventional linguistic point of view, we might say that it mixes up grammatical units and collocations. But that is like looking through the wrong end of a telescope. From the point of view of the speaker or writer, the selection of one word will constrain the choice of succeeding words. For example, given the choice of a noun, there is a relatively high probability that the next word will be a relative particle, and that this will be followed by an adjective; or to take individual words, given *orang*, that the next words will be *yang* followed by *lain*.

At the present stage, we have the technical means of counting the occurrence of strings of words or tags, what is rather more of a challenge is to know what to do with this information. Our parser takes a bottom-up approach to the formation of phrases and sentences, and takes only grammatical tags and individual word forms into account. This means that a frequent combination of words is treated exactly the same as a very rare one, and ignores the ready-formed phrases and expressions that make up a large part of written texts and a greater proportion of spoken texts. When human readers and listeners encounter expressions such as *negara yang lain* or *yang tidak*, they surely do not process them afresh each time and incorporate them into the grammatical tree.

Conclusion

In this paper we have reported some of the preliminary frequency figures obtained from the analysis of our corpus, and already some interesting facts about the language have begun to emerge. The use of classifier nouns, apparently a characteristic feature of Malay, actually turns out to be rather rare. The global figures almost certainly mask significant differences between genres. Our search includes about 50,000 words of newspaper articles, and it would be worth extending the search further to see to what extent in this kind of text they are used at all. The relative rarity of adjectives gives us an insight into what constitutes a typical noun phrase in Malay The impression given by introductory descriptions of Malay suggest that something like *rumah besar itu* is a typical noun phrase, but such an idea is contradicted by the fact that adjectives are not in fact particularly common in this position. These are things to follow up in future research.

We have so far limited our investigation to bigrams. The next stage is of course to look at trigrams, and longer n-grams. The limiting factor is the size of our corpus. A corpus of 2.3 million words is a good size for preliminary studies of the syntax, and the bigrams that emerged as frequent seem intuitively reasonable. Of course we cannot prove that the bigrams we found to be frequent really are frequent in the language as a whole; and we have to anticipate that some rankings reflect the composition of the corpus. The only way we can test and prove the validity of our rankings is by building and analysing a bigger corpus. We certainly need a bigger corpus to obtain useful results for longer n-grams. The other factor to take into account is the representativeness of our corpus. We need a far bigger database of processed texts from which we can select a representative sample before we can make substantiated claims about what word sequences are common, or indeed more generally what is normal in the Malay language.

The consideration of frequencies has important consequences for the design of an automatic parser for Malay Most parsing whether automatic or manual (for example in the drawing of phrase structure trees) is concerned with hierarchical structures in syntax, and with what is theoretically possible with the syntactic rules. However, since we have a database designed so that we can extract syntactic information from it, we can find out what structures are common and which are less common. For example, conventional grammars describing adverbial expressions of manner in Malay list the use of *dengan* with a kata sifat, or of *secara* followed by a kata sifat. Actually the first of these is rare and the second very rare. Adverbial expressions are most commonly formed with a simple kata sifat modifying the verb.

If we take a conventional hierarchical approach to syntax, frequency information might seem interesting but marginal in importance. From the point of view of the user of the language, on the other hand, it is central. When we compose sentences in writing or in speech, it is important to know what is likely to follow a given string of words or word classes, or the probability of some expression y given some expression x. The learner of Malay has to expect *sebuah* to be followed by the name of an object of a particular type, or that a noun will be followed by *yang* in cases where a relative clause might not be expected in English. At the very least, our current work is leading to an enriched description of the Malay language, in which grammatical possibilities are ranked in order of their frequency.

References

Abdullah Hasan. 1974. *The Morphology of Malay*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Bar Hillel, Y 1964. ed, *Language and Information*. Reading: Addison-Wesley

Knowles and Zuraidah (in press a) *Word class in Malay: corpus-based approach*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Knowles and Zuraidah. 2004. 'The notion of a "lemma": headwords, roots and lexical sets,' *International Journal of Corpus Linguistics* 9:1.

Muhadjir, Bobby A.A.Nazief, Mirna Adriani, Kiswartini Mangkudilaga & Multamia RMT Lauder. 1996. *Frekuensi Kosakata Bahasa Indonesia*. Depok: Fakultas Sastera Universitas Indonesia.