Open Journal of Modern Linguistics, 2014, 4, 728-738
Published Online December 2014 in SciRes. http://www.scirp.org/journal/ojml
http://dx.doi.org/10.4236/ojml.2014.45063

Scientific Research

Processing Malaysian Indigenous Languages: A Focus on Phonology and Grammar

Asmah Haji Omar

University of Malaya, Kuala Lumpur, Malaysia
Email: asmahomar@um.edu.my

Received 3 October 2014; revised 16 November 2014; accepted 4 December 2014

Abstract

Malaysian indigenous languages are of two entirely different families: Austronesian and Austroasiatic. The former consists of Malay and all the languages of Sabah and Sarawak, while the latter the aboriginal languages found only in Peninsular Malaysia. Except for Malay and a few more in Sabah and Sarawak, most of these languages have not been put into writing. This means that no writing system has been ascribed to them, despite the fact that quite a number have been described in terms of phonology, morphology and syntax. From the descriptions available, one gets a picture of their typologies and systems for processing purposes. Concerning typology, there is not much difference between the two families as far as phonemic inventories go, but there are differences in the phonological structures of the syllable and the word. As for morphology, the Austronesian languages are agglutinative, while the Austroasiatic ones are isolative. There is also a difference in the syntactical status of the word, where the former has the two categories of the full word and the particle, and the latter only the full word. This last mentioned difference leads to a divergence between them in the types of phrase, the clause, and the complex sentence. Natural language processing (NLP) is a methodology which is now being applied in the analysis of various aspects of languages. This paper discusses the constraints faced by most of the Malaysian indigenous languages in the application of this methodology.

Keywords

Austronesian, Austroasiatic, Natural Language Processing, Typology

1. Introduction

The concept of processing a language, and to be exact a natural language, is a new one. It came into being with the availability of the computer, which with a suitable software, is able to do so many things at the same time. It is a methodology of analyzing languages. The outcome of this methodology is equivalent to that analyzed

How to cite this paper: Omar, A. H. (2014). Processing Malaysian Indigenous Languages: A Focus on Phonology and Grammar. *Open Journal of Modern Linguistics, 4,* 728-738. http://dx.doi.org/10.4236/ojml.2014.45063

manually by linguists. This goes to show that analyzing language to determine various aspects of its systems and structures as well as its meaning has been going on since the beginning of the modern linguistics era, with the publication of Saussure's Cours de Linguistique Générale[1], or even before, but this age-old methodology has never been known as "processing". Indeed, the idea borne by the word "processing" is in accordance with the mechanics. As with other aspects of life, analyzing language has evolved from one that is labor intensive to one that is machine-assisted.

With language, before the computer takes over the job there is a long chain of action and conceptualization that has to be undertaken on the object, which is the language corpus. The corpus which contains the data to be analyzed has to be readable by the computer. This entails a higher degree of uniformity in the orthographic presentation of the corpus to be fed into the machine, compared to the traditional way adopted by linguists in treating a set of corpora collected from the field as a sacrosanct piece of information, one that cannot be interfered with, for the purpose of "cleaning up". In natural language processing (NLP), it is only after the corpus has been cleaned up can the process such as tagging be done at whatever level of language that serves the objective of the researcher.

This paper is about the state of the art of the processing of Malaysian indigenous languages. It will be seen that only three of these languages have the potential of undergoing NLP due to various constraints.

2. Malaysian Indigenous Languages

For its land size, consisting of a small peninsula in the west (known as the Malay Peninsula or Peninsular Malaysia), Sabah, Sarawak and Labuan (the latter three on Borneo island), Malaysia has a high density of indigenous languages. A rough estimate should be slightly over 100. What is meant by an indigenous language is one which has a stable community of speakers with a considerable time-depth, a genetic relationship with other native languages in the same geo-linguistic region, and a tradition of being recognized as a native language by the peoples themselves.[2]

Malaysian indigenous languages are divided according to any of the two language families or stocks. The first is the Austronesian family which represents the major group in terms of the number of languages (about 80), and with this the number of speakers. Members of this family are found in Peninsular Malaysia, Sabah, Sarawak, and Labuan. The Austroasiatic family, the minor group, is represented by some 20 languages and they are found only in Peninsular Malaysia. They are the languages of the aborigines, or *Orang Asli* (original people) (Omar, 1975, 1983). Among the aborigines, there are also groups that speak dialects of Malay (which is an Austronesian language) or languages closely related to Malay. Speakers of both the Austronesian and the Austroasiatic familes are known as *Bumiputera* (children of the soil).

The Bumiputera group consists of 62.8% of the total population of Malaysia of about 25 million people. This percentage may be taken to reflect the number of native speakers of Malaysian indigenous languages. Of the Bumiputera group, native speakers of Malay (inclusive those of all its dialects) comprise about 60% of the population of Malaysia. It goes without saying that Malay is the most dominant language of Malaysia in terms of the number of native speakers. At the same time its dominance is further manifested in the fact that it is the national and official language of the country, and hence is spoken by almost all, if not all, of the people of the land. As for indigenous languages other than Malay, both Austronesian and Austroasiatic, the number of native speakers form a gradient from several hundred thousand to less than a hundred (see Lewis, 2009).

3. The Malay Language

The centrality of Malay is not confined to its position as the common language medium in the primary domains in the life of Malaysians, as in governance and education, as well as in inter-ethnic communication. It has been the only lingua franca among peoples of the Southeast Asian islands, and the medium used by these peoples with foreigners. Malay is now the national and official language of three nations, namely Malaysia, Indonesia,

[1]Linguists have accepted the idea that modern linguistics began with the introduction of the general linguistic theory by the Swiss linguist, Ferdinand de Saussure, and this was taught to his students in his lectures between 1902 and 1911. His compendium of lectures was published posthumously by his students, giving the title *Cours de Linguistique Générale*, first published in 1915.

[2]The criterion of time-depth taken on its own appears to be rather tricky. The Portuguese language spoken by a small community of descendants of the Portuguese who ruled Melaka from 1511 C. E. to 1640 C. E. can now be said to have a time-depth of 900 years. In another 1000 years, its time-depth will be a millennium. But history has shown that it is not native to Malaysia, and this is further supported by the fact that it does not belong to the language families to which Malaysian indigenous languages belong.

and Brunei Darussalam. This means that there are three national varieties, and each of these supra-varieties has its own local dialects.

From centuries past, Malay has been about the only native language of the three countries mentioned above that has received a great deal of attention from scholars both within and outside the region, who have conducted research on various aspects of its systems and structures. Research on the language includes the reconstruction of older linguistic forms, giving explanation of linguistic change, and from there to the spread of the language in mainland Asia and the islands of Southeast Asia, and beyond.

The history of Malay as a written language began in the seventh century as evidenced by four stone inscriptions found in Southern Sumatera and the Bangka Island, which are now territories of Indonesia. These are texts written in the Pallava script of Southern India, using the high variety Malay of the period, narrating the success story of the ruler in his conquest of territories and in the spread of Buddhism (Omar, 2013: Chapter 3; G. Coedès 2009). These stone inscriptions appear to be the only evidence of the use of the Pallava scriptin writing the Malay language. It was Islam that came three centuries later that gave the Malays a permanent writing system in the form of the Arabic script which they have indigenized with the addition of new symbols to suit the phonological system of the language. In doing so, the early Malays gave an indigenous label to the script, the *Jawi* script. With this script the Malays were able to record their traditions in various genres, including those of the pre-Islamic ones.

When the Western powers came to the Malay world beginning with the Portuguese in 1511, followed by the Dutch in 1640, and then the British in 1786, another writing system was introduced for the writing of the Malay language. This is the Roman alphabet, known among Malaysians as the *Rumi script*, *Rumi* being a Malayized form of *Rome* from where the script originated. But this writing system did not gain much popularity with the Malays who preferred the Jawi writing. It was only when Malaya became independent from British rule in 1957 that the Rumi script was elevated to become the main writing system for Malay. This policy was undertaken for the purpose of modernization, as the Rumi script was closer to the world of knowledge, and to achievement in the fields of technology and the sciences. The Jawi script which previously had given valuable service to the Malays in the recording of their rich literary traditionsin the form of prose and poetry has been given second place in importance, and it is now mainly used in the teaching of Islam. What this means is that Malay texts before the second half of the 20th century are mostly in the Jawi script, while those printed after the period are almost all in the Jawi system of writing .

Malay lexicography has its beginning in the 15th century, with the arrival of traders and religious missionaries from outside the Malay world. The earliest known glossary is a Chinese-Malay word list, estimated to be of the date 1403-1511 C. E.[3] (Blagden & Edwards, 1930-1932). This was followed by an Italian-Malay word list, better known as the Pigafetta word list,as it was collected by Pigafetta, the Italian seaman who visited the Malay Archipelago in the early part of the 16th century (Pigafetta undated). There were many more word lists that emerged later on, mostly collected by travelers from the West, represented by the Dutch, the French, the Germans, and the English.

It can be said that interest from the outside world in the Malay Archipelago for various objectives such as trade, spread of Christianity, and colonization had generated a fertile growth of Malay lexicography which evolved from bilingual word lists to bilingual dictionaries. As shown by the Malay example, and most probably by a majority of languages of the world, dictionaries have their beginning as bilingual glossaries , illustrating the need of outsiders to know the language of a community other than their own. The monolingual dictionary for Malay made its debut only in the last decade of the 19th century.

Pigafetta's Italian-Malay word list seems to be the earliest evidence of the writing of Malay using the Rumi script. The graphemic renderings of Malay words in the list appear to be based on Pigafetta's perception of the Malay words he heard in his interaction with Malay-speaking peoples in the ports of the islands of the Malay Archipelago (Bausani, 1961). Similarly early dictionaries produced by Westerners as mentioned above give the spelling of Malay words according to the perception of the collector/compiler and is very much influenced by his native language. Arising from this practice, words in the Malay language were spelt in the Rumi script in many different ways. It was only in 1904 that a standard spelling system was formulated and implemented for use in the schools in Malaya and Singapore. This spelling is known as the Wilkinson spelling system, after the Chairman of the Spelling Committee, J. R. Wilkinson, who at that time was also the Director of Education of the

[3]The year 1403 C.E. was the year the first merchant fleet from China arrived in the Malay kingdom of Melaka, with the Muslim Admiral Cheng Ho as head of delegation. The first settlement of Chinese on Malay land took place circa the early part of the 15th century in Bukit Cina (Chinese Hill) in Melaka.

Federated Malay States and the Straits Settlements of Penang, Melaka and Singapore. Another spelling reform took place in the 1924 with minor changes to the Wilkinson system, to suit Malay phonology. The person responsible for these changes was a Malay language specialist in the Department of Education, by the name of Zainal Abidin bin Ahmad, better known by his pen-name Za'ba. This amended system came to be known as the Za'ba system as well as the school system, as it was used in the schools (Omar, 2008: p. 90; Nawang, 2005: p. 126).

With the implementation of the New Education Policy of 1970 which stresses the importance of Malay as the main medium of education from the primary to the tertiary level, there was a need to spell the language in such way that it should be able to take inneologisms in the form of academic and scientific terminologies which are sourced through English. There had to be a way of transferring them into the Malay language without much loss of their original spelling in the English language, but at the same time giving them a Malay phonological identity. As the national language of Indonesia, known as bahasa Indonesia, is basically the same language as that of Malaysia, and plus the fact that Indonesia also shared the same mission as Malaysia in wanting to elevate her national language to be a significant medium in the expression of scientific matters, a co-operation between the two countries was established to achieve this mission. The first product of this language cooperation was a common spelling system which was officially declared in 1972 as the official spelling for both countries. It is this system that is in use until today for the writing of Malay, not just in Malaysia, Indonesia, and Brunei Darussalam, but wherever Malay or bahasa Indonesia is used or taught (A full story of this language co-operation is given in Omar, 1979).

The presence and utilization of this common spelling system in the long history of writing Malay in the Rumi script, has given stability in the production of texts of all kinds. The following four properties have paved the way for Malay to undergo NLP:

1) Written texts in great quantity in various genres.
2) A standard spelling system in the Rumi script which makes it easier for computer-assisted processing compared to the Jawi script.
3) Full descriptions of its phonology and grammar.
4) Dictionaries in the monolingual and bilingual modes.

However, NLP of Malay has not really taken off on a significant scale. There are Ph. D. theses, though few, written by Malaysian students for the degree in local and foreign universities, and these are based on data from specific corpora.

The Dewan Bahasa danPustaka, or DBP (which in English is known as Institute of Language and Literature), was set up by the government in 1956 on the eve of the Malayan independence, with the responsibility of developing and promoting the Malay language as national and official language as well as the main language of education. One of the tasks of this institute is to produce dictionaries for Malay. It has done so for both the monolingual and the bilingual types. Its monolingual dictionary the *Kamus Dewan*, which has now reached its fifth edition, is the foremost authoritative reference in Malay lexicography. Likewise are the DBP's bilingual Malay-English (*Kamus Melayu—Inggeris Dewan*), and English-Malay (*Kamus Inggeris—Melayu Dewan*) dictionaries. All these dictionaries were first compiled in the days when Malaysia was not yet acquainted with NLP (Omar, 2008).

The DBP has since built a data base for Malay, to be used for various purposes in the promotion of the Malay language. Uppermost in its list of objectives is the compilation of what I would call a "mother dictionary", which will be the source or the referent point in the compiling of other dictionaries for Malay. MIMOS (Malaysian Institution of Microelectronic Systems) which was established in 1985, and placed under the Ministry of Science, Technology and Innovation, has also started to include NLP in its programme of activities.

4. Other Indigenous Languages: State of the Art

Of indigenous languages other than Malay, the only two languages that can be said to be ready for NLP are Iban of Sarawak, and Kadazandusun of Sabah. They have the properties (i)-(iv) which have been ascribed to Malay, though not in terms of quantity. The compendium of written corpora in each of these two languages is still thin compared to Malay. Both these languages were first given a writing system, in the Rumi script in the late 19th century, for the purpose of translating religious texts from the Bible, from the English source. The translation was done by English-speaking Christian missionaries with the help of native speakers. Christian mission schools were set up in the villages where most of these ethnic groups were domiciled (Banker, 1984, Department of Education Sarawak 1984, KCA Sabah, 1989). Up to this point in time, these two languages were still languages

of memory, a term first used by Swaan[4], i.e. languages which had not been given a written form.

The production of non-religious texts came much later, with the establishment of the Borneo Literature Bureau (BLB) by the British colonial government in 1952. The main activity of the BLB was collecting and publishing folk traditions of the ethnic groups of Sabah and Sarawak. All these traditions had hardly been put in writing prior to that date. Probably because the office of the BLB was in Kuching, Sarawak, the bulk of folk traditions in the form of folk tales, native belief systems, customs, traditions etc. published by BLB were those of the Iban (formerly known as Sea Dayak) people. Publications of folk traditions of other groups in Sarawak and in Sabah were too few for mention, including those of the Kadazan and Dusun people of the latter state.

As mentioned earlier in this paper, the rendering of texts of folk traditions in the native languages were done according to individual perceptions of the speech sounds of the narrators, which were then transcribed into the Rumi script. There was no standard form of spelling for the language of the texts collected. And this was in tandem with a situation where a standard form of speech was non-existent. For example, Iban texts published by the BLB between 1952 and 1977[5] are spelt in many different ways. To a field linguist, like myself, variations in the spelling of words may represent a variety of dialects. Even though the transcription may not be accurate, this was a starting point for the field linguist to probe into the various dialectal variations. But at the same time I could see from these Iban publications some semblance of a uniformity in the spelling of words in texts which were collected by speakers originating from one and the same region. For example, texts collected by Iban speakers from the Second Division of Sarawak appear to be similar in the spelling of Iban words, and these may differ from those compiled in the Third Division by dialect speakers of this region. It was when Iban became an important mother tongue[6] subject in the national schools in Sarawakin 1987 that efforts were undertaken to formulate a single standard spelling system for use in the schools, as well as in the publishing of Iban books, newsletters, and pamphlets (Department of Education Sarawak, 1964; Dayak Cultural Foundation, 1995; Bahagian Pembangunan Kurikulum, 2007).

Kadazandusun of Sabahbecame a mother tongue subject taught in national schools in the state in 1997, ten years after Iban. Like Iban, there were already texts written in the Kadazandusun language, but in the early days of its appearance in the Rumi script the spelling was the result of the perception of the writers/collectors of folk traditions. In the 1990s, knowing that the chance to have their language taught in the national school, educators and leaders in the Kadazandusun community were already preparing for the admission of their language into the national school system, and the first task undertaken was formulating a standard spelling system. This activity went hand in hand with the writing of texts for teaching in the schools (KCA Sabah, 1989; Lasimbang, 1998).

Having a standard orthography for a language means that there is already a phonological description of this language, and this was the case of both Iban and Kadazandusun. The descriptions rendered for these two languages may not be comprehensive for a linguist. However, the identification of phonemes and their variations should be sufficient to decide on corresponding graphemes, using the symbols of the Rumi script. In the transferring of oral texts to writing, there is another requirement, and that is the identification of language units at higher levels above the phoneme. These are the morpheme, the word, and the sentence. This task, as it were, has to a certain extent been accomplished for both Iban and Kadazandusun (see Omar, 2013; Omar & Sandai, 2012a, 2012b; Lasimbang, 1994, 1998).

Dictionaries are a help in NLP. Of all these other indigenous languages, Iban appears to have more than any other, even compared to Kadazandusun (Lasimbang, 1994). These are mostly bilingual Iban-English dictionaries (Richards, 1981; Sutlive & Sutlive, 1994). So far there is only one Iban-Malay dictionary, and an Iban monolingual dictionary (Tun Jugah Foundation, 2011).

As for indigenous languages other than Iban and Kadazandusun, most of them are still a long way away in the production of corpora for NLP. This means they do not have all of the properties (i)-(iv) posited above. The great majority of them are still languages of memory, in terms of phonological and grammatical descriptions, while some others have been given basic descriptions sufficient to write down oral texts. Even if some basic descriptions do exist, there is a lack of written corpora. An example is the Mah Meri language, an Austroasiatic

[4]The term “language of memory” was first used by Dutch sociologist, Abram de Swaan, to refer to languages which have not been given a writing mode. See de Swaan, 2001, 2003.

[5]In 1977, BLB was taken over by the Dewan Bahasa danPustaka which turned it into its Sarawak Branch, which is extant to this day (see Omar, 2008).

[6]In Malaysia a mother tongue of any ethnic group that is taught in the national schools is known by the official term “Pupils’ Own Language” or POL for short. A POL class can be set up when at least 15 pupils (i.e. their parents) ask for it.

language spoken in Sepang, a district not far from the Kuala Lumpur International Airport in Peninsular Malaysia. There is already a sufficiently comprehensive description of the phonology and grammar, and also a sufficiently comprehensive glossary in trilingual mode of Mah Meri-Malay-English consisting of lexical items of Mah Meri life, traditions and environment (see Omar (Ed.), 2014). However, there is hardly any text written in the language.

For the fieldworker to get lists of vocabulary of a language unknown to him is not a difficult task, as there are various techniques in eliciting words of different domains. The same goes with eliciting types of sentences. However, to have a written corpus in these languages, one has to source from their traditions. In other words, the researcher has to get the informants to tell their stories. From my own experience in the field among the "small language" communities in Sabah and Sarawak, as well among the Orang Asli in Peninsular Malaysia, it has not been smooth going in getting informants to tell stories in their own language. I had no problem with the Iban in the longhouses upstream in my early days on doing research on their language, although at that time I was just learning to pick up their speech system through staying with them in their longhouse. With informants of most of these so-called small languages, they prefer to tell their stories in the lingua franca they share with the researcher, that is Malay, or specifically the Malay dialect used by their community members when speaking with people of other ethnic groups. Most of the time they would say that they have no story to tell, or that they have forgotten all those tales of their ancestors (Omar, 2014a; Jan et al., 2014).

I attribute this reluctance of theirs to a misapprehension they have of their language, thinking that it is an inferior speech system compared to others, and associated with this is a pretension of memory loss. This misapprehension is brought about by the presence of a powerful lingua franca, Malay, which is also the national language. To illustrate, when my team of researchers and I were working on the MahMeri language in 2001 through to 2002, the informants, men and women, were reluctant to provide a lengthy discourse on any topic using their language. But they were ready to do it in Malay, and we managed to have a collection of their verse forms, songs, folktales, and incantations in the medium of the Malay language. After months of establishing an acquaintance with them, we managed to coax their chieftain (*TokBatin*) to relate to us the well-known story of *Si Tenggang*, a son who dismissed his mother as a poor village woman when he took home a princess from abroad for his bride. For this unbecoming behavior, he received aretribution from the power up above in which he, his bride and retinue, and the whole ship which took him home, was turned into stone. This metamorphosis in the form of a limestone hill is supposed to be the Batu Caves (meaning "caves of stones") in Kuala Lumpur, which has become a popular tourist attraction.

However in certain parts of Sarawak, to my knowledge, Malay or a local dialect of Malay is not the only lingua franca. There is an equally powerful one, and that is Iban. My own survey of the Bintulu area in 2005 showed that Iban was much more spoken in the market places and in the town area compared to Malay. There was a Tatau longhouse not far upriver from the Bintulu town area, occupied by members of two different languages which were in the endangered list, the Tatau language having only 75 speakers, and Lugat with 30. The traditional ceremony welcoming me was conducted in Iban, which fortunately I could understand. When I asked them to tell me in their own language how they came to be settled in that longhouse, so that I could record their story in their native tongue, the story came out in Iban with Malay code-switching (Omar & Ghazali, 2014).

A study on language choice in the multiethnic and multilingual district of Sarikei in Sarawak by Mohammed Azlan Mis shows that while each of the four ethnic groups of Malays, Iban, Melanau and Chinese choose their own language in the home domain, the favorite speech systems in intergroup communication are Sarawak Malay and Iban. Due to the habit of using these two languages in such a situation, there has emerged a mixed code of Sarawak Malay and Iban (Mohammed Azlan Mis, 2011).

We could take another approach in our acquisition of corpora in an indigenous language by having the native speakers themselves to collect them from members of their own community, in the same way the BLB used community members to record their community's stories using the medium of their own language. With the Mah Merimy team and I couldn't find native speakers to co-operate in carrying out this task, let alone the youths who were literate as they were busy joining the rat race in a competitive world. Most of the time they couldn't understand why we were interested in their language.

With today's technology one can still conduct NLP using recorded speech of a particular language. But without a written corpus to fall back on, and plus the fact that there does not exist a description of its phonology and grammar, this may prove to be a very difficult task even with the assistance of a native speaker who is no doubt an excellent speaker in the language. Being a good speaker is not the same as being able to answer the linguist's

question when it comes to the tagging of grammatical unit, or the parsing of short stretches of speech into component units. The linguist will still have to carry out this task on his own. All this goes to show that the description of the phonology and grammar is basic to the processing of these indigenous languages. Phonology provides the basis for the identification of grammatical units, as well as to the formulation of a spelling system.

5. Typology of Malaysian Indigenous Languages

In the introduction to this paper I have mentioned that Malaysian indigenous languages belong to two entirely different families. The implication is that each family is characterized by its own typology. Identifying the typology of a language at any level is useful for NLP. For this, one can start with the premise that the typology of any particular language within a single family is a reflection of the typology of other members within the same family. On this premise, and with evidence from analyses that have so far been conducted on languages of both the Austronesian and the Austroasiatic families of Malaysia, one can get a general picture of the typologies of these two families. As those who have worked on language typology know, when we say that a language with typology X at the phonological or morphological level, it means that the general picture given by this language is that of type X at this level of analysis. At the same time it does not mean that this language does not have type Y as its phonological or morphological characteristic; only that type Y is not significant in terms of number and frequency compared to X.

Discussion on the typology of the languages below is given according to the levels of language analysis: phonological and grammatical. Only the types which are indigenous are taken into account in this typology.

6. Typology of phonological Units

At the level of the phonology, there appears to be a great deal of similarity in the types of phonological units between the two languages, as given in **Table 1**.

An obvious difference between the two families of languages is in the syllable structure, where the Austroasiatic languages admit consonant clusters, up to three consonants, as onsets of the syllable; hence, the indigenous syllable structure of this group can be written as (C)(C)(C)V(C).

On the other hand, the indigenous syllable structure of the Austronesian languages is (C)V(C), an indication that consonant clusters do not feature as on sets of syllables. If they do, there are two possibilities that may explain the existence of this feature. One is due to an internal phonological change, showing the contraction of the first syllable of a disyllabic word to a cluster, resulting in a monosyllabic word. The formula for this type of change is as follows:

Syllable 1 - *Syllable* 2

#VN - XY#

V represents any vowel

N any nasal consonant

X any non-nasal consonant homorganic with N

Y any element(s) occurring after Y

- silence before and after the word

Deletion occurs on the V of the first syllable, as given below, causing a reduction of a two-syllable word to become a monosyllabic one, and the latter takes a homorganic nasal-oral cluster as onset:

#VN - XY#→ #NXY#

Examples: anjaŋ → njaŋ (Malay: third uncle or aunt)

anda?→ nda? (Malay: fourth uncle or aunt)

ənti → nti (Iban: if)

ənday → nday (Iban: no, not)

Table 1. Types of phonological units.

Phonological Units	Austronesian Languages	Austroasiatic Languages
Vowel Inventory	Between 6 and 8 All cardinal vowels	Between 6 and 8 All cardinal vowels
Consonant Inventory	Between 16 and 19 With phonemic and prosodic **ʔ** and **h**	Between 16 and 19 With phonemic and prosodic **ʔ** and **h**
Syllable Structure	(C) V (C)	(C) (C) (C) V (C)
Word Structure	Disyllabic	Disyllabic

The second explanation as to the occurrence of consonant clusters as syllable onsets in Austronesian languages may be seen in the many neologisms from English that have entered the Malay language. Even with some change in their graphemic representations, these neologisms still retain the clusters. Examples are words such as *proses* (process), *klinik* (clinic), planet (*planet*), *struktur* (structure), *psikologi* (psychology), etc. This particular phonological feature has been accepted as part of the phonological structure of Malay. Slowly they are making their entrance into Iban and Kadazan with the transfer of the more general items in the vocabulary from Malay, such as *proses*, *projek*, and *klinik*.

Clusters in the Austronesian languages are mostly of the type nasal-oral on condition these two elements are homorganic. And the position of this type of cluster is inter-syllabic, i.e. with a juncture between them, with the nasal ending the first syllable, and the oral homorganic consonant beginning the second syllable. Hence, in syllable division one can place a hyphen between the two. A cluster of this type is known as abutting cluster, a term first popularized by Abercrombie (1967). Examples:

Malay /kampoŋ/ : /kam-poŋ/ (village)
/nanti/ : /nan-ti/ (wait)
/gaŋgu/ : /gaŋ-gu/ (disturb)

Iban/kampoŋ/ : /kam-poŋ/ (forest)
/dindiŋ/ : /din-diŋ/ (wall, partition)

TambunanDusun /lintun/ : /lin-tun/ (descend)
/noŋgo/ : /noŋ-go/ (where)

As seen from the examples given above, abutting clusters occur in the word-medial position whereas the non-abutting ones are as onsets of syllables.

In terms of indigenous feature, no consonant cluster occurs in the word-final position. With the development of Malay as a medium in teaching the sciences and technology in schools and universities comes the acceptance of English terms with word-final consonant clusters, such as *kompleks* (complex), *eksport* (export), *kloroform* (chloroform), etc.

7. Typology of Grammatical Units

A comparison of the typologies of grammatical units of both families of languages under comparison is given in **Table 2**.

Main differences between the Austronesian and the Austroasiatic families can be located at four main levels: morphology, word, phrase and clause. At the level of morphology, word- forms appear to be more complex in the former compared to the latter. Complexity in the latter is seen only in the presence of reduplication, but this morphological type is that of full reduplication. In the Austronesian languages, there are three types of reduplication: (i) of the first syllable of the root word; (ii) of the root of the complex word; (iii) of the whole word.

The higher complexity of word-forms in the Austronesian group is also seen in the use of affixes. Of all the

Table 2. Types of grammatical units.

Grammatical Units	Austronesian Languages	Austroasiatic Languages
Morphology	(i) Agglutinative: affixes(prefix, infix, suffix) Reduplication: first syllable of root word, root of complex word, whole word	Isolative Reduplication: whole word
Word	Full words Particles	Full words
Phrase	Head-Modifier (H-M) Prepositional	Head-Modifier (H-M)
Clause	Independent Dependent	Independent
Sentence	Subject-Verb-Object (SVO)	Subject-Verb-Object (SVO)
Complex Sentence	Paratactic—with & without conjunctions Hypotactic—with conjunctions	Paratactic—without conjunctions

three types of affixes given in **Table 2**, the one that is most dominant is the prefix, that is to say prefixes occur in all the languages in the group, some tending to have more of this type of affix than others. For example, Malay has a total of 30 single unit affixes, consisting of 16 prefixes, five infixes, three suffixes, and six split affixes (Omar, 2014b). On the other hand, Iban has 12 affixes, 11 of which are prefixes, and the other a suffix (Omar, 2013). Tambunan Dusun, spoken in Sabah and which has not been given a written tradition, has20 verbal affixes, 16 of which are prefixes (Omar, 1983: pp. 214-243).

What is meant by a single unit affix is an affix which consists of a single bound morpheme. In the Austronesian languages these single unit affixes are prefixes, infixes, suffixes, and split affixes. In **Table 3** below, data from three languages are given as a comparison in the division of these affixes into their different types. As will be seen, there is clear evidence that Austronesian languages are typologically prefixal. **Table 3** does not include complex affixation where the lexico-grammatical derivation of certain categories of words requires the use of two or three affixes. This type of agglutinative complexity is more prevalent in Malay than in any of the other languages.

As shown in **Table 2**, another point of divergence between the two groups of languages under comparison isthe grammatical type(s) of words each has. The Austronesian languages have both the full words and the particles, whereas the Austroasiatic languages only have the former type. If particles appear in the latter, they are transfers from a language of the Austronesian family, and this other language is usually Malay (see Omar (Ed.), 2014, all chapters on word classes and phrases).

Particles comprise prepositions and conjunctions. Due to the absence of these two grammatical subcategories, the Austroasiatic languages do not have prepositional phrases which in the Austronesian languages function as adverbial phrases of place, time, direction, comparison etc. Concepts borne by such phrases in the Austroasiatic languages are conveyed by full words. Similarly, the absence of conjunctions in Austroasiatic languages also means that these languages do not have dependent (or subordinate) clauses. This in turn leads to them not having complex sentences of the hypotactic type. In the Austronesian languages hypotactic complex sentences are marked by the presence of subordinating conjunctions.

As for the paratactic type of complex sentence, in the Austronesian languages these are formed by combining an independent clause with at least one other independent clause, with or without a coordinating conjunction. When no such conjunction is used, the paratactic relationship between the clauses can be identified in the intonation contour which pitches down at the end of the preceding clause before the one following it is uttered. In written language, this is realized in the use of the graphic symbol, the coma (,). Paratactic complex sentences in the Austroasiatic group are of the same type as the Austronesian paratactic sentences which do not make use of a conjunction

As for the typology of the simple sentences, both groups of languages are of the SVO type. In conversation and story-telling another structure appears to be as significant, and this is the Theme-Rheme (or Topic-Comment) structure. Written Iban tends to use more and more of the SVO type of structure, whereas Iban texts collected by the BLB in 1950s and 1960s show a greater usage of the Theme-Rheme structure. The tendency towards SVO is due to the influence of Malay which is the national and official language, and the learning of this language in the school.

Table 3. A Comparison of affixes in Malay, Iban, Dusun Tambunan.

Affixes	Malay	Iban	Tambunan Dusun
Prefix	**16** (6 verb, 4 noun, 3 numeral, 3 adjective)	**11** (7 verb, 2 noun, 1 adjective, 1 numeral)	**17** (15 verb, 1 noun, 1 adjective)
Infix	**5**(4 noun, 1 adjective)	-	**4** (verb)
Suffix	**3**(2 verb, 1 noun)	**1**(verb)	**1** (noun)
Split Affix	**6**(2 verb, 3 noun, 1 numeral**)**	-	**4** (3 noun, 1 verb)
Total	**30**	**12**	**26**

8. Conclusion

A natural language has to have the four properties specified in this paper for it to be suited for NLP. Among the many Malaysian indigenous languages, only Malay, Iban and Kadazanduusun can be considered suited for this type of processing. It has also been shown in this paper that descriptions of phonology and grammar, as well as a sufficient amount of corpora are needed for this method of language analysis. Most of the Malaysian indigenous languages still have a long way to go before NLP can be applied on them.

References

Banker, J. E. (1984). *The Kadazan/Dusun Language*. In J. W. King, & J. K. King (Eds.), *Languages of Sabah: A Survey Report* (pp. 297-324). Canberra: The Australian National University.

Blagden, C. O., & Edwards, E. D. (1930-1932). A Chinese Vocabulary of Malacca Malay Words and Phrases Collected between A.D. 1403 and 1511 (?). *Bulletin of the School of Oriental and African Studies, 6*, 363-397.

Dayak Cultural Foundation (1995). *Atur Sepil Jaku Iban* (*The Spelling System of Iban*). Kuching: Dayak Cultural Foundation.

de Saussure, F. (1960). *Course in General Linguistics*. London: Peter Owen Limited.

de Swaan, A. (2001). *Words of the World: The Global Language System*. Cambridge: Polity Press.

de Swaan, A. (2003). Asia's Affairs with English, Hindi, Filipino and Malay. In A. H. Omar (Ed.), *The Genius of Malay Civilisation* (pp. 365-411). Tanjong Malim: Institute of Malay Civilisation, Universiti Pendidikan Sultan Idris.

Department of Education Sarawak (1964). *The Full Teaching Syllabus for Junior Secondary Schools*. Kuching: Department of Education Sarawak.

Jan, J. M., Zaid, A. R. M., & Shamsudin, K. (2014). Antologi Cerita, Pantun, Lagu, dan Jampi Bahasa Mah Meri. *Jurnal Bahasa Jendela Alam, 8*, 9-61.

KCA Sabah (1989). *KOISAAN Language Symposium: Towards Standardisation of the Kadazan Dialects*. Souvenir Book. Kundasang, 13-15 January 1989.

Kurikulum, B. P. (2007). *Sistem Jaku Iban di Sekula* (*Iban Speech Systems for Schools*). Kuala Lumpur: Ministry of Education Malaysia.

Lasimbang, R. (1994). Kadazan Dusun—Malay—English Dictionary. *Proceedings of the Third Biennial International Conference of the Borneo Research Council*, Pontianak, 10-14 July 1994.

Lasimbang, R. (1998). Kadazandusun Mother Tongue Education. In K. K. Soong (Ed.), *Mother Tongue Education of Malaysia Ethnic Minorities* (pp. 96-99). Kuala Lumpur: Dong Jiao Zong Higher Learning Centre.

Lewis, M. P. (Ed.) (2009). *Ethnologue: Languages of the World* (16th ed.). Dallas, Texas: SIL International. http://www.ethnologue.com/16

Nawang, A. H. (2005). *Memoir Za'ba*. Tanjong Malim: Penerbit Universiti Sultan Idris.

Omar, A. H. (1975). The Verb in Kentakbong. In *Essays on Malaysian Linguistics* (Chapter 19). Kuala Lumpur: Dewan Bahasa dan Pustaka.

Omar, A. H. (1979). *Language Planning for Unity and Efficiency: A Study of the Language and Status Planning of Malaysia* (Chapters 9-11). Kuala Lumpur: Penerbit Universiti Malaya.

Omar, A. H. (1983). *The Malay Peoples of Malaysia and Their Languages*. Kuala Lumpur: Dewan Bahasa danPustaka.

Omar, A. H. (2008). *Ensiklopedia Bahasa Melayu*. Kuala Lumpur: Dewan Bahasa danPustaka.

Omar, A. H. (2013a). *Sejarah Ringkas Bahasa Melayu*. Kuala Lumpur: Department of Museums Malaysia.

Omar, A. H. (2013b). *The Iban Language of Sarawak: A Grammatical Description* (Second and Enlarged Edition of 1981). Kuala Lumpur: Dewan Bahasa danPustaka.

Omar, A. H. (2014a). Analysis of Folktales, Verse Forms and Incantations of the MahMeri. *Jurnal Bahasa Jendela Alam, 8,* 1-8.

Omar, A. H. (2014b). *Nahu Melayu Mutakhir*. Edisi Kelima. Kuala Lumpur: Dewan Bahasa danPustaka.

Omar, A. H. (Ed.) (2014). *The Mah Meri Language: An Introduction*. Kuala Lumpur: University of Malaya Press.

Omar, A. H., & Ghazali, K. (Ed.) (2014). Indigenous Minorities of Bintulu: A Sociolinguistic Mapping. In A. H. Omar, & N. Norahim (Eds.), *Linguistic Minorities: Their Existence and Identity within Larger Communities* (Chapter 6). Kuching: Universiti Malaysia Sarawak.

Omar, A. H., & Sandai, R. (2012a). *Fonologi Bahasa Iban: Fonologi Jaku Iban*. Tanjung Malim: Penerbit Universiti Pendidikan Sultan Idris.

Omar, A. H., & Sandai, R. (2012b). *Morfologi Bahasa Iban: Morfologi Jaku Iban*. Tanjung Malim: Penerbit Universiti Pendidikan Sultan Idris.

Richards, A. (1981). *An Iban-English Dictionary*. London: Oxford University Press.

Sutlive, V., & Sutlive, J. (1994). *Handy Reference Dictionary of Iban and English*. Kuching: Tun Jugah Foundation.

Tun Jugah Foundation (2011). *Bup Sereba Reti Jaku Iban* (*An Iban Monolingual Dictionary*). Kuching: Tun Jugah Foundation.