CHAPTER 4

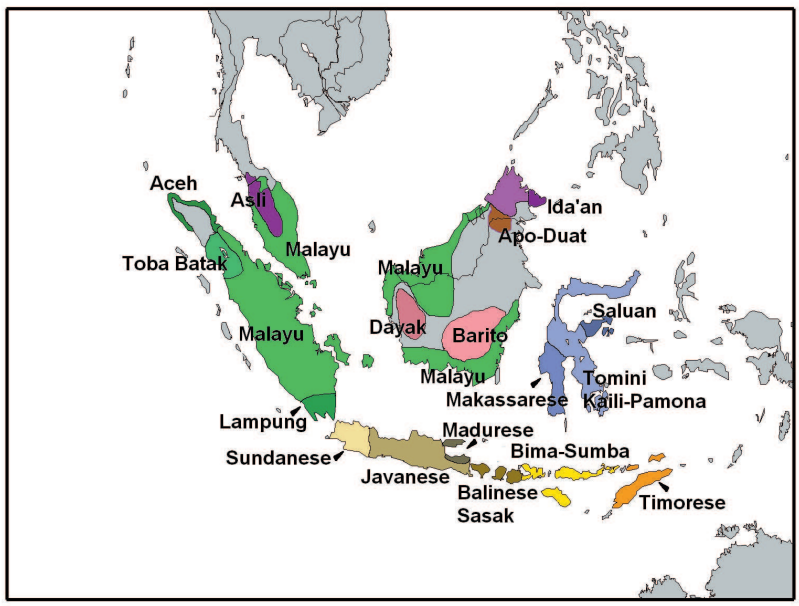
The Languages of South-East Asia

Bahasa Indonesia
Bahasa Melayu
Thai

4. **The languages of South-East Asia**

4.1. **The Austronesian languages**

The Austronesian languages are spoken in Malaysia, Singapore, the Indonesian Archipelago and the Philippines. Malaysia, Singapore, Indonesia and the Philippines have standardized their languages. For Malaysia and Singapore this language is called "Bahasa Melayu" (Standard Malayan) and for Indonesia "Bahasa Indonesia" (Standard Indonesian). These languages are characterized by affix modifiers of words. Some are prefixes or suffixes separately attached to words, other affixes consist of both a prefix and a suffix. Infixes occur too, but are treated as a separate lemma in the dictionaries.



The Austronesian languages spoken in Malaysia, Singapore and Indonesia. The lingua franca, the standardized languages Bahasa Melayu and Bahasa Indonesia, gain influence at the expense of local languages.
[2]*Inspired by De Grote Taalatlas (The Atlas of Languages).*

In Bahasa Indonesia affix modifiers vary in complexity (see the sample presented below):

-i	me-
ber-	me-....-i
ber-....-an	me-....-kan
keber-....-an	menge-...-kan
se-....-an	berse-...-an
menye-....-i	menye-...-kan

These prefixes are clearly existent in Bahasa Indonesia, but also in Bahasa Melayu, and in the local languages. In the Balinese language the Bahasa Indonesian word *bangun* (rise) is identical to the Balinese word, but the affixes attached to the word are written differently, e.g. *berbangun* (rising up) becomes *ba-bangun-an* in Balinese. Despite the different forms of affixes, the principles of the word modifiers are the same. In Tagalog, one of the main languages spoken in the Philippines, the -an affix modifies *ibigy* to *ibigy-an* (shall give to somebody to be named thereafter).

Another characteristic of the Austronesian languages is the reduplication of words.

> Bahasa Indonesia
> Nouns: kupu-kupu (butterflies),
> Pronouns: saya-saya (the reduplication of I),
> Adjectives: bagus-bagus (beautiful),
> Verbs: duduk-duduk (sit about),
> Numbers: satu-satu (one)

Affixes in reduplication encapsulate the reduplication: *besar* in *membesarkan* is reduplicated to *membesar-besarkan*. Affixation is applied to the unit having a distinct meaning. An artificial mechanism which would generate words as *membesarkan-membesarkan*. Such a mechanism would erroneously accept combinations that don't exist in real language. In other words non-sense would be approved. Some spellcheckers use these artificial mechanisms to camouflage their inability to build proper dictionaries.

The nature of affixing makes the Austronesian languages very distinct from the Indo-European languages and neighbouring Austroasian languages. The European languages French, Spanish, Italian, and Greek also use prefixes but typical of these languages are the modulation of the verb ending to express time in an ongoing stream of the syllables. Contrary to this ongoing stream of syllables

the Austronesian affixes are invariable and additional affixes can be added to further change the semantics of a lemma. Usually lemmas are short; one or two syllables. These short lemmas expanded by affixes create a rather irregular structure of syllables.

4.1.1. **Hyphenation**

As discussed earlier, American hyphenator designs often use a theory that syllables are equally distributed throughout the word. This is valid for the French word *fu~tu~ro~lo~gue* but not not for the Bahasa Indonesian word *peng~a~dil~~an* (the court) or *se~per~ang~kat~an* (a complete couple). It is just on the boundaries of the affixes that the linear hyphenation model is applied falsely. The effect of falsely applying these concepts results in an increase of errors at the affix boundaries. There are a few ambiguities too: *meng~u~kur* or *me~ngu~~kur*, *ter~a~ngan(-a~ngan)* and *(ber~te~rang-)te~rang~an*, etc. The hyphenator does not hyphenate ambiguous syllables.

The *TALO model is not linear but starts with a language model that describes the characteristics of a language, in this case, the characteristics of the Austronesian languages, in particular Bahasa Indonesia and Bahasa Melayu. This model reduces the complexity by choosing linguistic units that belong to the languages themselves, instead of using incorrect assumptions about the language's nature.

Bahasa Indonesia and Bahasa Melayu are intimately related to each other. Governmental commissions have even accepted standards between the two languages (e.g. the conference with Brunei, Malaysia and Indonesia, 1991). They use the same orthography.

Therefore a unified hyphenator structure is feasible; just one hyphenator engine which is highly accurate, even usable for most of the local languages.

4.1.2. **Spelling**

For Bahasa Indonesia and Bahasa Melayu the linguistic structures are similar, but spelling focusses on the differences between the neighbours. The inner mechanisms of the speller engine do not need to be different, but the dictionaries are specific for each language. We have built a first series of dictionaries for Bahasa Indonesia and Bahasa Melayu that will confirm the orthography of the main Bahasa Indonesia and Bahasa Melayu dictionaries[3,4]. They will contin-

ue to expand as time goes by. Re-spelling capabilities finally will instantaneously know erroneously spelled words and apply correction automatically.

The Bahasa Indonesia and Bahasa Melayu orthographies have standardized spelling of European words[7]. Still English or Dutch orthography quietly enters documents: *expansive* instead of *ekspansif*, or *stabiel* instead of *stabil*. These are the cases to be corrected automatically. During the development of the lexicons this type of erroneous usage has been put into data bases. Spelling errors which cross the word boundary can be included too:

penanggungjawab (should be *penanggung jawab*)
pertanggung jawaban (should be *pertanggungjawaban*)
proklamasi republik Indonesia (should be *Proklamasi Republik Indonesia*)
24,500 orang (should be *24.500 orang*)

Spelling documents include punctuation checks too. Here, preferences between Basaha Indonesia and Bahasa Melayu may differ. The Republic of Indonesia has inherited many of the Dutch regulations while the Republic of Malaysia was strongly influenced by the British empire. Both nations use a different decimal system, respectively the European continental and the Anglo-Saxon system (Rp2.477.946,00 or Rp2,477,946.00).

Broadening our view we named the hyphenator "Euro Asia Hyphenator" and the speller ''Euro Asia Speller". For Quark XPress it became the Hyphenator XT and the Speller XT, for Adobe's InDesign it became Smart Hyphen and Smart Speller.

4.1.3. **Acknowledgement**

We thank Mr. Lim Bun Chai of the Kompas Daily for the fruitful discussions and support which have been stimulating the development of the Bahasa Indonesia hyphenator.

4.2. **The Thai language**

The Thai language belongs to the Tai family of languages. All members of this group are located in South-East Asia. Lao, of Laos, and the Chang language of northern Burma also belong to the Tai family of languages.
Thai is spoken by about 40 million people. Thai has its own script, introduced by king Ramkamhaeng in 1283. The script has its origine in India. The Thai alphabet consists of more sounds than European languages. There are 44 consonants and most vowels are not represented by an individual letter but by a mark written above, below, before, or after a consonant, pretty much creating a syllable script[8,9,10]. The Thai language is a tone language with the diacritics of the script indicating middle, low, falling, high or rising tone marks. The script runs from left to right; majuscules don't exist nor do punctuation characters. However, the main difference between Thai and most other languages is the way sentences are written: that is WITHOUT spaces between the words. Together with compounding this is one of the major obstacles in printing.

4.2.1. **Compounding**

In the Thai language compounding is a principle to create new words having a different meaning than the meaning of their original components[10].

"to understand เข้าใจ" is derived from "to enter เข้า" and "heart/spirit/mind ใจ"
"train รถไฟ" is derived from "vehicle/car รถ" and "fire ไฟ"
"electricity ไฟฟ้า" is derived from "fire ไฟ" and "sky ฟ้า"
"lightning ฟ้าแลบ" is derived from "sky ฟ้า" and "pain แลบ"

The meaning of a compound always implies more than just the combination of the meanings of its components. A married couple for instance is more than a man plus a woman

4.2.2. **Sentences, Words, and Syllables**

A Thai sentence is a single unit, words are not separated from each other by blanks. To divide Thai sentences into words is not fundamentally different from hyphenating European words into syllables.

An English sentence written as a Thai-like sentence would appear as:

"theflowersofthefinestgreenhousesarenotwasted"

When this sentence is divided into individual words the context will be the decisive factor. A "green house" (i.e. a house painted green) is quite different from a

"greenhouse" (i.e. a glass building for growing vegetables or flowers), but the context makes clear that this sentence is about greenhouses.
Other sentences could be rather conflicting: "Godisnowhere" could be divided into "God is nowhere" or "God is now here". This phenomenon is not very different from pecularities found in European languages, e.g. the English language: a "rec-ord" (i.e. report, document) or a "re-cord" (i.e. maximum achievement)!
The Thai people read words in context.

A last example of a transcribed Thai sentence and a word-by-word translation (actually the Thai write sentences without spaces)[11]:

mi:	sa:mi:	phanraja:	ramruaj	khu:	nung	maj	mi:	lu:k
is	husband	wife	rich	couple	one	not	have	child

If the word 'ramruaj' ร่ำรวย is divided into 'ram' ร่ำ and 'ruaj' รวย the word 'ram' could be placed at the end of the sentence:

mi:	sa:mi:	phanraja:	ram

However, that would change the meaning of the sentence. 'Ram' means "to scent" (spraying perfume), so the meaning of this text line would be changed to: "is husband wife spraying perfume".
This is absolutely not allowed.

4.2.3. **A second layer for in-word hyphenation for newspapers**

Newspapers require narrow columns. Therefore newspapers do wish to hyphenate Thai words, but only at boundaries that can not be misinterpreted. The Thai Hyphenator consists of two layers, the first layer divides sentences into words, the second layer divides words into syllables. The division of words is a separate procedure distinguishable from the optional division in syllables of individual words.
The Thai word for "chairman", a compound with the *Sanskrit* prefix *pra*, can be divided as: ประ^ธานี.
The word date (day of the month or year) วันที่ can not be split into วัน and ที่, because the meaning would become day followed by 1) place, 2) in , those, 3) that, plus the other words in the sentence. Despite this limitation a high density of hyphenation can be realized thanks to *TALŌ's two-layer technology of the Thai language model.

4.2.4. **Acknowledgement**

We thank Ms. Pariya Huntjes-Suwannaphoom for her successful efforts in building a learning corpus of the Thai language, we also thank her for the fruitful discussions and support which stimulated the development of the Thai sentence segmentation program.

4.3. **References**

1 Barber, C.C. Dictionary of Balinese-English (Vol. 1 & 2), Aberdeen University, Occasional Publications, Aberdeen, 1979.

2 De Grote Taalatlas (the atlas of the languages), Schuyt & Co, Haarlem, 1998.

3 Kamus besar Bahasa Indonesia, Departemen Pendidikan dan Kebudayaan, balai Pustaka, Jakarta, 1999 & 2001.

4 Kamus Malaysiana, Ensimal (m) Sdn. Bhd. Kuala Lumpur, Malaysia, Edisi Pertama, 1994.

5 Sneddon, J.N., Indonesian, a comprehensive grammar, Routledge, London and New York, 1996.

6 Teeuw. A., Indonesisch-Nederlands woordenboek, KITLV Publisher, Leiden, 1996.

7 KAMA/zz., A(C)-Pedoman Umum Istilah B.M.-May 2000(anum).

8 Woordenboek Thai-Nederlands, L.J.M. van Moergestel, Nangsue, Zaandam, 1995.

9 Woordenboek Nederlands-Thai, L.J.M. van Moergestel, Nangsue, Zaandam, 1995.

10 Thai-English Student's Dictionary, Mary R.Haas, Stanford University Press, Stanford, California, 1964.

11 The structure of Thai Narrative. Somsonge Burusphat, the University of Texas, Arlington, 1991.