



I. Proudfoot

Concordances and classical Malay

In: Bijdragen tot de Taal-, Land- en Volkenkunde 147 (1991), no: 1, Leiden, 74-95

This PDF-file was downloaded from <http://www.kitlv-journals.nl>

I. PROUDFOOT

CONCORDANCES AND CLASSICAL MALAY

A concordance is a particular way of displaying the form of a text. It is 'any arrangement of the words in a text in which the occurrences of the words are alphabetized (or ordered according to some principle) and in which some of the environment of the word is presented along with each occurrence' (Smith 1969: xxi). An index, on the other hand, gives a selective guide to significant information in the text. The index is a companion to the text, a commentary on the text. The concordance is not interpretive in this sense. It rearranges the surface form of the text, providing a basis upon which some interpretation can be more conveniently built.

Compiling a concordance is not an intellectual activity. It is a task of immense clerical drudgery. The great obstacle to concordance-making had always been the prodigious amount of tedious labour it required. The first biblical concordance, of the Vulgate, was made in the thirteenth century by a team of 500 clerics. Without such resources, the making of a concordance could be the life-work of an individual. It would consume thousands of hours of the time of pettifogging scholars and the volunteers they employed to transcribe words and contexts on to slips of paper for later indexing and sorting. Understandably, under such circumstances, concordance-making was restricted to texts of the greatest cultural significance.

The advent of the electronic computer has therefore been a boon to the would-be concordance-maker. It is easy enough to instruct the electronic computer to act as a clerical drudge, and it will do so accurately, swiftly, and tirelessly. While computer concordance-making has been feasible for the last 30 years¹, recent developments have made the technology widely accessible. Two software packages, one developed by Oxford University, the other by Brigham Young University, now put basic concordance-making in the hands of microcomputer users (*Micro-OCP* 1987, *Word-Cruncher* 1989). This makes computer time effectively free. Additionally,

¹ The first computer-generated concordances appeared in 1957. For a hair-raising description of concordance-making on the punch-carded valve-driven Univac of that day, see Parrish 1962.

I. PROUDFOOT is a lecturer in Asian History at the Australian National University, where he also obtained his Ph.D. Previous publications include *Ahimsā and a Mahābhārata story*, Canberra, 1987, and 'A formative period in Malay book publishing', *Journal of the Malaysian Branch of the Royal Asiatic Society* 59, 1986. Dr. Proudfoot may be contacted at the Faculty of Asian Studies, Australian National University, P.O. Box 4, Canberra 2601, ACT, Australia.

the tiresome task of preparing the text upon which the concordance will be based has been greatly lightened by the advent of the optical character reader. Initial data entry may no longer be necessary. Checking and editing, though, remain burdensome enough.

Ironically, the readiness with which concordances can be made today has not brought unalloyed joy to practitioners of the art. Indeed it is the cause of some anguish. Already in 1970, Bessinger was worrying: 'Since today's computers are so unwontedly powerful and can perform routine operations so much faster than human beings can, they may save scholars vast amounts of physical labour and perhaps even a little intellectual labour. The only real question is, can this power of theirs be used economically? Can we afford it?' (Cameron et al. 1970: 4-5).

It is now worthwhile concordancing a text with appeal to a limited scholarly audience. Has the currency thereby been debased? If it is now possible to generate massive output on demand, how is this to be published? To keep concordances to manageable size, should they be made selective? To do so ignores the power of the new technology. It might have been unbearably tedious for a clerk to concord 'and' in the English Bible, but for a computer the task is trivial; indeed it is more troublesome to exclude 'and' from the concordance than to include it. But even a selective concordance is voluminous. If its appeal is to a small specialist audience, who will publish it? If it is not published, the making of the concordance was not useful. Although the computer has also reduced the costs of preparing the printed edition, the print medium remains expensive. Is the answer to publish not in print, but in microfilm, or to record the results digitally, on disk or tape, for (re-)retrieval by computer?

The problem is a misapplication of technology. The computer is being used to treat the text as a clerk would. This is no surprise, for the first applications of new technology are almost always solutions to old problems. For concordance-making, the immediate result is a bottleneck when it comes to storing and physically reproducing torrents of clerical output. Using the super-clerk this way fails to take proper account of its speed and cheapness.

A better approach has been applied in the field of bibliography. An early use of the computer in bibliography was the listing of titles in the KWIC style. The KWIC ('key word in context') index is a modified concordance of titles using only predetermined key words. Indexes of this kind will be familiar to readers of this article from *Dissertation Abstracts International*, or (in attenuated form) from the subject index of the *Social Sciences Citation Index*. These bibliographical tools assumed ever greater bulk, until their recent replacement with data-base searches on-line or on CD-ROM disks. This solves the problem of publishing voluminous concordance output. Information is produced only on demand to satisfy specific inquiries. But this new approach is not just a cheaper way of publishing prepared output. As information is gathered anew for each inquiry, both the ele-

ments of the inquiry and the display of results can be tailored to specific needs. It is no longer necessary to compromise with one organizing principle – say, the alphabetic listing of key words – which is judged most likely to satisfy the average user.

Library card catalogues, which allowed only a few points of entry to the collection, have gone the same way. The catalogue card is fast yielding to the on-line database.

The day of the concordance printed in book form will, also, soon be past. When a text can be searched and the results tabulated almost instantaneously, it is no longer rational to preserve the results of one simple search in print or any other medium. The future lies with interactive investigation of the text with specific needs in mind: perhaps to list all contexts in which a word occurs, but also to list collocations of a word with others, or conjunctions of a word with other features of the vocabulary and syntax of the text, or the distribution of a word within the text, etc. The concordance may be a 'general-purpose working tool for the study of literature' (Howard-Hill 1979: 4), but who will use a bus when a taxi is on hand? Already one micro-computer program, *WordCruncher*, will run complex searches on text and generate elementary statistical information interactively.

So, for the time being, we are briefly in the lurch between an old clerical tradition and a new age of interactive text analysis. Until more powerful techniques for interactive text analysis become widely accessible, and likely users more comfortable with it, the printed concordance still has a role.

Uses of printed concordances

While the limitations of the traditional concordance must be appreciated, it is easy to be too dismissive of the contribution it can make to a field like Malay studies. It may be a defective tool, and one whose day will soon have passed, but just now it has much to offer.

The main uses of the concordance flow from the principle adopted for the arrangement of its entries, under key *words*. Because the words of the text are also units of sense, the concordance can be used for an index, albeit an unreflective and indiscriminate one. It is a poor substitute for an analytical index, but the conjunction of words and contexts at least allows the user to import a little of the indexer's analytical judgement. In practice, it has been as a comprehensive index of words that the concordance has been most used and most valued. Biblical concordances, the earliest and most used, have especially had this role.

The presentation of words in context makes the concordance particularly well adapted to the study of the meaning of words. Indeed a major step towards systematic philology is the compilation of a concordance. This has been true of Biblical and classical scholarship in the West, Quranic scholarship in the Islamic world, and Vedic (and now Epic)

scholarship in India. The world's great historical dictionaries have been built up from slips or cards recording apparently important words in context in dated sources. The raw materials of such a dictionary thus resemble nothing so closely as an array of concordances (though the dictionary slips are gathered selectively and may record a wider context than is warranted in a concordance). Conversely, a concordance is a ready-made pile of dictionary slips. The availability of concordances thus greatly assists the compilation of dictionaries, whether based on historical principles or frequency of usage.

In the realm of classical Malay this work has hardly begun. Dictionaries useful for classical Malay (Klinkert 1885; Wilkinson 1901-03, 1932) include citations from texts, though the selection is not systematic, nor arranged on historical principles (which would have been difficult in any case, as most texts used were manuscripts dating from the nineteenth century). A foretaste of the rich discoveries which can be expected is Matheson's (1979) intriguing study of the changing meaning of '*melayu*' in southern court histories. This involved a great investment of time simply in finding where the word was used, which could have been short-circuited if concordances of the texts had been available. (Though it should be noted that the study depends upon a fuller and more sensitive appreciation of context than a concordance alone would provide.) What can be done with '*melayu*' can be done with any number of culturally or linguistically interesting terms. Clusters of near-synonyms may be particularly revealing: inter alia Matheson's work suggests further consideration of '*tanah*', '*kerajaan*', '*daerah*', '*alam*', etc. This is still a virgin field. Light may fall on meanings, dialects, and genres.

Analogously, a concordance may throw light on small-scale syntactic structures. A fine example of a study of this kind - made without the help of a concordance - is Tol's (1984) analysis of *suruh* constructions in seventeenth-century Malay. What knowledge have we of the use of even the commonest conjunctions and enclitics in classical Malay? Accumulation of such knowledge makes the concordance a useful aid in resolving cruxes in a text edition. In this way, the making of concordances may play a dynamic part in the editorial process. 'Now one also needs a concordance from which to prepare an edition from which to make a concordance', comments Bessinger drily (Cameron et al. 1970:9).

Concordances are not only about words (or phrases). They are also about the relationships between words (or phrases). The first concordance was indeed intended, as the name suggests, to demonstrate agreements (*concordantiae*) in the text of the Vulgate Bible. Like its successors, it was concerned therefore with the form of the text, in the belief that in the form lay part of the text's meaning. In this respect, the concordance is a re-arrangement of the text designed to illuminate some aspects of its form. It will allow inferences about choice of vocabulary and elementary patterns of collocation; it could conceivably contribute to questions of

authorship² – though it is not particularly suited to any of these inquiries.

The brief contexts of the concordance do promise, however, to throw some light upon formulaic composition at the level of the phrase. Parry's first generalization of the techniques of oral bardic composition to literary works was facilitated by concordances of Homer. This experience is relevant to classical Malay. Sweeney (1987: 73-76) has pointed to formulaic conventions in the manuscript tradition, and my study (Proudfoot 1967: ii 193-197) of the prose mousedeer texts threw a little light on scribal use of formulaic parallelism. Concordances promise to enhance our perception of formulae and variations at the phrase level, and perhaps point indirectly to formulaic structures on a larger scale. The concordance-based study by Koster and Maier of *Syair Ken Tambuhan* (1982; cf. Akehurst 1981: 156-157) explores some of these possibilities.

In general the concordance is an apt tool for lexicology, but for most other linguistic studies and for stylistics it is a blunt instrument. For other purposes, other methods of calculating and displaying patterns in the text are more fruitful (Hockey 1980: 79-143). In the medium term, the greatest benefit of today's concordance-making may prove to be the provision of computer-readable texts which can be exploited more flexibly than was feasible or imaginable in the days of the pen and ink. For the time being, though, I expect that the use to which concordances of classical Malay texts will most often be put will not differ from the most frequent use of other concordances: they will be used as comprehensive indexes to texts hitherto unindexed or selectively indexed.³

Policy questions about the form of a Malay concordance

The day of the computer-generated concordance may be brief. But it is with us now. If we will make a concordance of a classical Malay text, we must decide the most useful shape for its unidimensional display. This involves taking positions on issues which have perennially divided concordance-makers, whether they have wielded quill or qwerty. The questions are:

Should the concordance be lemmatized?

Should its spelling be normalized?

Should it distinguish homographs?

How should context be delimited?

and

Should all words in the text be included?

For diplomatists, the concordance is a mechanical rearrangement of

² Is *Hikayat Hang Tuah* the work of two hands? Kassim Ahmad's hypothesis (1964: xii-xiii) could not be disproved, but might be confirmed by careful study of frequent vocabulary usage. Such studies, however, lie more in the domain of statistical analysis.

³ The brief indexes available are generally to proper names or intuitively selected 'unusual' words. Even the more substantial (Matheson and Andaya 1982; De Josselin de Jong 1961) continue to focus on such personal names and toponyms rather than topics.

aspects of the surface form of the text; analysts accept that some editorial interpretation may enhance the utility of the concordance as a reference work. These questions will therefore be answered differently by supporters of diplomatic and analytical practice (Hart 1979: 230). The answers will also be influenced by the particular characteristics of the material being treated.⁴ To date these issues have been debated mainly by scholars working with European languages, with alphabetic writing systems, and dealing with corpora of drama, verse, or punctuated prose. We, though, must consider the particular needs of classical Malay.

Should it be lemmatized?

Concordance entries are almost always listed according to the order of key words. Should the key word be the form which actually appears in the text? Or should it be the 'dictionary' form of the word, the 'lemma'. There is feeling in the concordance world both for and against lemmatization. Stances have been adopted on both practical and principled grounds.

Lemmatizing the words of a text is far harder than making a concordance from them. Lemmatizing inevitably means taking the trouble to resolve ambiguities and irregularities. To take a Malay example, *beribu* would have to be construed as *ber+ibu* or *be(r)+ribu* according to context. Although it is conceivably within the capacity of the computer to resolve such ambiguities, such sophisticated programming is beyond present invention. The requirements of context sensitivity would in the end be similar to those for machine translation, which is still an unachieved ideal. In practice, then, lemmatization may mean extensive editorial intervention. The diplomatist sees two detriments here. First, the processing of the concordance is considerably slowed. Without lemmatization, a concordance can be generated swiftly and entirely mechanically. Second, as human judgement is involved, pure mechanical predictability is sacrificed to some degree. The more interventionist position regards the concordance as more than merely a re-arrangement of the undigested text. The text has already been changed by subjecting it to a new principle of organization: why should the restructuring not go further?

A secondary objection to lemmatization raised by the diplomatists is that the 'abandonment of the simple alphabetical arrangement of headings - to which all users have equally easy access . . . necessarily restricts the general usefulness of the concordance' (Howard-Hill 1979: 4).⁵ Failure to lemmatize may not be very painful in most Indo-European languages for, apart from the occasional aorist or passive past participle prefix and some vowel modulation, the Indo-European conjugations and declensions rely

⁴ On all these issues, see Hockey's excellent chapter on 'Word Indexes, Concordances and Dictionaries' (1980: 41-78); also Fogel 1962.

⁵ Howard-Hill's policy is exemplified in the prestigious *Oxford Shakespeare Concordances* (Howard-Hill 1969-72). His stance is rebutted convincingly by Lusignan 1980; it is taken to extremes by Fleury 1986: 240.

on suffixes. Consequently even without lemmatization, the grammatical variants and derived stems of a root tend to cluster together in the alphabetic sequence of key words. Even so, with highly inflected languages blessed with a well established lexicographical tradition, lemmatization seems both natural and desirable (cf. Burton 1982:200). Fleury, for instance, felt lemmatization to be self-evidently indicated for any Latin concordance (1986:240).

What holds for Latin applies *a fortiori* to Malay, where common prefixes would scatter verbal and derived forms up and down the alphabet. It seems unproductive to allow lists of the occurrences of *alami*, *dialami*, *kualami*, *mengalami*, *pengalaman* etc. to be separated by dozens or hundreds of pages. Nor will lemmatization make access to Malay forms less easy. Rather the contrary: all who are accustomed to using Malay dictionaries are thoroughly inured to lemmatization.

It may be that for certain studies of morphology and syntax, and for some studies of formulaic patterning, an unlemmatized concordance of raw words would be preferable, but for the purpose to which the concordance is best adapted, lexicology, and if the concordance is to be used as a topical index, lemmatization is clearly desirable. A further argument for lemmatization of Malay is that the presence of affixation is a socio-linguistic variable marking register (Benjamin 1988). The play of this feature is both better observed and better accommodated in a lemmatized listing.

Should its spelling be normalized?

As in other traditions before print, the spelling of classical Malay manuscripts is notoriously inconsistent. Standardization of spelling is a product of print and mass education. In the transition from manuscript to print culture, a standard form is also more characteristic of learned languages, and less characteristic of vernaculars. Compare the spelling of Latin and English in Shakespeare's England, or of Arabic and Malay in *kitab* texts. The written dialect we call classical Malay falls toward the middle of the spectrum running from learned language to vernacular. It is a learned language of literature and religion while other Malay dialects are (for some of its users) mother tongue or lingua franca. In addition, the poorly adapted Arabic script used to write classical Malay is the source of some spelling variation, principally in vowelizing; though at the same time that script's skeletal nature keeps at bay some spelling variation which a fully phonemic script might have revealed. It is clear, too, that spelling conventions vary with time and place (*besar* : *besyar*, etc.). Some spelling practices may reflect dialectal differences (*pula* : *pulak*, *memunuh* : *membunuh*, etc.). These possibilities are often remarked by editors of Malay texts, but the variations found in their manuscripts are rarely reported consistently

in their text editions.⁶ This is regrettable. Working in the more elaborated scribal tradition of Java, Behrend (1987:xiv-xv, 362-367) points out that spelling and handwriting styles may illumine the geographic and social provenance of texts – or what he calls the ecology of transmission.

Confronted with manuscript variation of this kind, European concordance-makers have been divided over the desirability and usefulness of normalizing spellings in a concordance. The problem is to accommodate pre-print variability within a print-conceived structure (Ong 1982:101, 123-126). The diplomatic position is that a concordance should not emend the text upon which it is based. With a printed text 'even typographical errors should also be left as they are'.⁷ In this way no feature of the text will be lost. But by the same token, the concordance is no longer based on *words*, but on graphemes. Coverage of a word will be dispersed according to its different graphic forms, difficult to locate without a multitude of cross-references – which the editor must add. The concordance thus loses potency as a survey tool for those interested in words (Gärtner 1980; Burton 1982:200) – or in lemmatized words. However, it is possible to retain the word-based organization of the concordance without sacrificing spelling variation by normalizing the key word headings which define the order of entries in the concordance while retaining graphic forms in the illustrative contexts. This strategy serves both analytic and diplomatic interests efficiently (Parrish 1962:10).

But what in classical Malay do we mean by the words and forms of the text? Until the early twentieth century, a scholarly edition of a classical Malay text might use the Arabic script. In modern Arabic, Persian, and Urdu studies this would pass unremarked, and concordances naturally follow suit. But during the twentieth century, in Indonesia and Malaysia, Arabic script has been largely displaced by Roman spelling. Modern editions of classical Malay texts are now published in Roman script, both to make the text accessible to modern readers, and because the Roman spellings are more complete phonemically. To be useful, a concordance must be based on an accessible published text. Concordances of classical

⁶ It is customary to devote a section of the editor's introduction to cursory notes on the 'spelling of the text'. Alternatives to this practice are providing a parallel version in Arabic script, as did Jones in his edition of *Hikayat Sultan Ibrahim* (1983), or providing the edited text in Arabic script, as did Shellabear in his edition of *Hikayat Seri Rama* (1915). Shellabear used Arabic script not only to preserve antique spellings, but also because in Singapore in 1915 the Latin script was widely used only for 'low' Malay.

On attitudes to dialectal or historical divergence from *a priori* standard form, see Teeuw 1959: 152-154.

⁷ Hockey 1980: 65, adding 'Corrections to them can be inserted in brackets in the text, but the original should not be deleted'. But even the very diplomatic *Oxford Shakespeare Concordances* emend rank typographical errors.

Malay texts will therefore be based on Roman transcriptions of manuscript material.⁸

The implications are twofold. The spelling of a Romanized text has already been significantly normalized. Interesting graphic variation in the manuscript has been filtered out of most editions in the process of transliterating from Arabic to Roman script. On the other hand, a variety of modern transcriptions has been employed. Not only have the official systems of spelling and word division been reformed, but scholars have improvised upon them as well.⁹ There is, therefore, little value in retaining the spelling of the printed editions insofar as it is a matter of one transcription system or another. It is not interesting to know whether Mulyadi (1983) made *Indraputra* or *Indeputera* the hero of her text, nor that Skinner should have transcribed the same word as *Tjé'* in 1963, *Ché'* in 1965 and *Cik* in 1985. But it may prove interesting to know where or how often Mulyadi's text has *manusia* and *manusyia*. Such manuscript variations, conveyed through transcription, are worth retaining. Thus, nothing is lost in modernizing a transcription system, so long as it is done systematically so that any recorded variation is preserved. There is no cost in normalizing *tjahaja* to *cahaya*, or *tjaja* to *caya*, so long as the distinction between the variant forms is retained, with the occurrences of both interleaved in a common entry under the normalized headword, say *cahaya*.

Compounds present a sticky problem for the word-based concordance, as they raise the problem of fixing word boundaries. Fortunately the problem is only severe for compounding languages, a category which does not include classical Malay. Nevertheless, a few awkwardnesses will have to be dealt with. In the manuscript tradition, word division was partly indicated by allographic form but otherwise usually ignored. Where it is indicated, it is little more consistent than the spelling. Should *barangkali* be one word while *barang siapa* is two? This follows modern practice, but means that *barangkali* will be listed in the concordance after *barang* and not at all under *kali*, whereas *barang siapa* will be represented first among all the uses of *barang* and again under *siapa*. Or is it preferable to follow modern practice with the title *Yang Dipertuan*, which in a concordance will be buried under *yang* and the verbal forms of *tuan*, or is this form better listed separately as one word *Yangdipertuan*, which is not at odds with the manuscript practice? Cross-referencing is a common way of dealing with such difficulties.

At a lower level is the problem of defining words and affixes. In Romanized text editions more word divisions have been introduced than the manuscript form would support. In manuscript, for instance, *yang* is regularly joined to what follows. The spelling reform agreed to by the Indone-

⁸ An edition in Jawi with vocalized headwords (or Romanized headwords) is conceivable, though pointless while Jawi reading skills are not widespread. On this question, too, see Behrend 1987: x-xv.

⁹ Recently, Brakel 1975:95 on spelling and 1975:43 on word division.

sian and Malaysian governments in 1972 has moved even further from the manuscript tradition by separating *di* and *ke* from following words when they indicate spatial relationships. Romanized texts also distinguish the suffix *-kan* and *kan* as the abbreviated form of *akan*, although these are represented by the same manuscript form. Despite such grey areas, as the concordance must be based on published Romanized texts which give no access to manuscript word divisions, and as the headings of the concordance should be standard, predictable, and accessible, we are constrained to follow the modern conventions.

Should it distinguish homographs?

Homographs abound in Malay written in the Arabic script. A major task of the editor of a classical Malay text is to manage the distinguishing of homographs by vocalization in the Roman script. With Roman transcription, which is more complete phonemically, the limited number of homographs which remain are almost always homophones. In discussing Malay concordances made from Romanized text editions, therefore, the two terms are effectively interchangeable.

The diplomatic view is that homographs should not be distinguished as a matter of principle: 'it is best to let the machine do unaided as much as you can of the big job, to interfere with it as little as possible, to resolve that a machine concordance is different in kind from a manual concordance, and that this is not necessarily a bad thing' (Bessinger in Cameron et al. 1970:37; de Tollenaere 1976:123). By deciding to lemmatize the text, we have already moved away from this stance. The example of editorial intervention in lemmatization given above involved precisely distinguishing homophones (*beribu*). However, it may be unwise to go beyond what is necessary to support lemmatization. As a matter of practical utility, homophones are easily distinguished in their concordance contexts, while conversely it is difficult for an editor to detect them all before the concordance is compiled (Hockey 1980:63).

There are sound reasons for not trying to make distinctions. If the decision turns on etymology only, it may be a scholarly artifice. It is conceivable, too, that homophones may be used for deliberate ambiguity. But above all, at the present stage of Malay philology, decisions on what are varieties of sense or grammatical function and what are homophones may not always be clear. The high-frequency homophones which have bedevilled European concordance-makers are distinguished by grammatical function. Examples are *le* or *la* in French, *die* in German, or *that* in English. Among common analogous forms in Malay would be *yang* and *akan*. The distinction of the senses of *that* in English rests on a highly developed and conventionally accepted description of parts of speech the like of which has not begun to emerge in Malay studies (Benjamin 1988: 30-31). It is therefore premature to distinguish occurrences of *yang* and *akan* on functional grounds. Nor does it seem wise to distinguish homo-

phones by imposing *a priori* semantic distinctions; this tends to prejudge issues best resolved within the framework provided by the concordance listing. So long as an editor's decisions cannot be confidently predicted by the users of the concordance, the accessibility of the concordance will be degraded.

The costs of not distinguishing are slight. Statistics of vocabulary range and frequency will be marginally distorted.

How is context to be delimited?

An objection to many published concordances has been that the context in which words are placed has been limited arbitrarily for reasons of mechanical convenience. Some obloquy has been directed toward the KWIC format because its common form restricts context to one line of printed output. In the early years of computing, this amounted to no more than a few words. But the objection is not only that the context may be too brief, but also that it is not defined by the sense units of the text (while, of course, in hand-made concordances contexts can be crafted case by case).¹⁰ The question is, how are the appropriate sense units to be determined. For most European concordance-makers, this has been a manageable problem. If the text being treated is in verse, is a play, or is punctuated prose, then strategies for fixing relevant contexts are easily engineered (Spevack 1973: 18; Fleury 1986: 241). Classical Malay *syair* verses mostly observe the sense division of the couplet; so, in dealing with verse, a single couplet, or pair of verses for words at verse-end, is an aptly defined context. Classical Malay prose is almost wholly unpunctuated, and there is yet little understanding of how its natural sense units are delimited or interrelated (Sweeney 1987: 236-237).¹¹ The punctuation introduced into transliterated texts by the editors is largely intuitive. A delineation of context using hierarchies of punctuation, feasible for European texts, thus loses its attraction in classical Malay. Lacking alternatives, it is difficult to object to the arbitrary delimitation of context for prose text.

How generous should an arbitrarily defined context be? The greater the context reproduced, the more convenient the concordance becomes as a research tool in its own right. But for a text of any length, the price of generous contexts is spectacular bulk. The happiest compromise for most purposes is, I believe, the conventional KWIC format. With one 'key word in context' per line, the key words being arranged alphabetically in a column near the centre of the page, and as much context to left and right as can be managed, the KWIC format is economical. Moreover, as occurrences of the key word are aligned in a central column, the concordance

¹⁰ Thus Fleury wishes to remove useless words encumbering the text, at the expense of shorter contexts (1986: 241).

¹¹ It would be too optimistic to believe that Becker 1979 has advanced our understanding.

will display recurring patterns in the context quite effectively.¹² (To aid this purpose, prose and rhyming verse contexts should be distinguished, though the choice may sometimes be unclear in classical Malay texts.)

Should all words in the text be included?

Ideally, the concordance should be complete; in practice, many are not. The reason is entirely practical. Even with the spare KWIC format, the concordance of a text of any length quickly becomes unmanageable. For a printed book with, say, 200 pages of 450 words per page, even a tightly formatted KWIC concordance will likely run to 1250 pages. It is possible to reduce the bulk significantly by omitting the relatively very few words which occur with very high frequency (Fogel 1962:24).¹³ For example, the length of a concordance of *Tuhfat al-Nafis* could be reduced by one-third with the omission of its 20 most common words. These common words are typically conjunctions and prepositions; in the *Tuhfat al-Nafis* they would be (in order of frequency): *maka, itu, yang, raja, dan, ke, di, pun, dipertuan, dengan, ia, orang, muda, serta, dalam, baginda, sultan, Riau, apabila, syahdan*.

Concern for the world's forests and the ire of bookbinders may be good reasons for giving up access to *maka, itu, yang* and their ilk. Most concordance-makers have been persuaded. To more idealistic souls, such compromise is ignoble. In a veiled attack on the *Oxford Shakespeare Concordances*, Spevack (1973: 17-18) argues for completeness at any cost. His most telling point is that it is fallacious to assume that the common is unimportant.¹⁴ Common words can unlock both overt and latent characteristics of the text. Displaying conjunctions in context can illumine syntax, for instance; formulaic phrases hinge on common words; and the particles of negation or emphasis may be critical to style (Fogel 1962: 24; Bessinger 1969: xi, xvi). Further, with Malay philology still in its infancy, even the commonplace is interesting.

¹² Bessinger 1969: xvii; Burton 1981: 147; Tebben 1977: v. Sorting the entries of a particular word in order of their left or (usually) right context is implied.

There appears to be a scholarly aversion to the KWIC format which is not wholly rational. Howard-Hill, for instance, assumes that KWIC context will always be as limited as early line-printer output made it, and denigrates the value of sorting by context using a hypothetical example which trivializes the issue (Howard-Hill 1979: 50-55). Compare Burton's (1982: 205) real example of sorting by context. One suspects that the KWIC format is not valued because it does not try to emulate the dignity of the handmade typeset concordance: cf. Oakman 1979:79.

¹³ By omitting 131 common words in the Revised Standard Version of the Bible, Ellison was able to reduce the size of his concordance by 59% (Fogel 1962:24). With the *Tuhfat al-Nafis*, the reduction would be 62%.

¹⁴ His first argument that completeness runs with the grain of the new technology, wherein omission is actually more difficult than mindless completeness, is the argument of a zealot. In any case the concordance is to be published with an old technology: on paper. Note that Spevack does not approve of KWIC displays.

On the big issues of concordance-making, then, the positions which will prove most profitable for classical Malay are neither partisanly diplomatic nor analytic. The optimal approach might be labelled 'conservatively analytical'. Its method may be distilled into some working rules for concordance-making in Malay.

The point of departure is that a concordance can only be as good as the published edition upon which it is based; and if it is not based on the best published edition, then its value is diminished.

In form, the concordance:

1. should be ordered by lemma (dictionary-form key word);
2. should group derived forms of the lemma in a logical order under the lemma heading;
3. should list occurrences of the same word in order of the context which follows;
4. should use a standard spelling for the lemma headings;
5. should preserve graphic variants indicated by the edited text;
6. should distinguish verse from prose;
7. should give a complete treatment of all words, including the most frequent;
8. should illustrate words in an arbitrary context as generously as possible.

By following these guidelines, a useful body of lemmatized texts and concordances could be built up.

Making Malay concordances¹⁵

Implementation of the guidelines listed above involves two major tasks: the lemmatization of the text upon which the concordance is based, and the conversion of the lemmatized text into a well structured and formatted concordance.

Lemmatization

To create a concordance based on dictionary forms, the lemmatized form of every word in the text must first be determined. The lemmatization of a classical Malay text by hand would be a daunting process. To examine the text word-by-word and to tag the affixed forms and roots would be both time-consuming and error-prone. It would also prove highly repetitive, as the same or related forms kept reappearing. How many times might the

¹⁵ A pilot project was funded by the Australian National University in 1988. Programs for lemmatizing, sorting and formatting have been written in the Pascal and Snobol languages to run on the Macintosh personal computer. The first experiments have been conducted with permission of the KITLV on Mulyadi's edition of *Hikayat Indraputra*, 1983 (a copy of which concordance, covering liv + 914 pp., has meanwhile been deposited also in the KITLV library), and with the author's permission on Matheson's edition of *Tuhfat al-Nafis* (in progress).

long-suffering editor come across yet another *menyembah* needing to be tagged as having the *me-* prefix and the root *sembah*? One way of simplifying the process would be to make a preliminary concordance of the raw words of the text, which would bring together all the *di-* forms, for instance, for more convenient examination and tagging. But what of *-kan* forms? or *ke-...-an* forms? etc.

The Gordian knot may be cut by capitalizing on some features of classical Malay. The classical Malay manuscript tradition is expressed in a rather standardized dialect with a transparent morphemic system (Teeuw 1959:152-4; Benjamin 1988:15-16). Moreover, both the prose and verse styles of classical Malay are quite repetitive, being adapted to aural consumption. Repeated standard forms suggest a role for the electronic drudge. It is feasible to write an efficient program for computer-assisted lemmatization, which will pass through the text analysing and tagging the morphemic structure of each word. Such a program may be made smart by introducing elaborate rules for morphemic analysis, but a point of diminishing returns is soon reached. Editorial oversight will always be necessary, because it is not possible to know beforehand that the peculiar forms of every text have been foreseen, and because it is not efficient (nor feasible) to ask the computer to deal with occasionally ambiguous analyses when these can be readily resolved by an operator. In practice, the low diversity of vocabulary of classical Malay material makes it quite feasible for the operator to inspect every analysis the program proposes. In the *Tuhfat al-Nafis*, for example, a text of 149,000 words ('tokens') involves only 7,000 discrete word-forms ('types'). The aim, then, is a program which will offer its best guess as the proposed morphemic analysis. The smarter it is, the less often the operator will have to do anything other than signify assent. The program can be instructed to remember the operator's previous decisions and to apply them automatically if appropriate, or to incorporate them in formulating its later best guesses.

Experience suggests that a program operating on a fairly simple set of rules gets better than 9 out of 10 of its best guesses right. Occasionally the operator has to tell the program that the base of *memuat* is not *puat* (its best guess) but *muat*, but should the program later meet the form *memuatkan*, it will get that right. The memory of such decisions can be stored in a dictionary file built up by the program, so that even the need to make best guesses becomes infrequent. The decisions made in analysing one text can then be carried over to subsequent exercises.

The program creates a lemmatized text by tagging the unprocessed text. The requirements of a tagging system are that it:

1. should indicate morphemic construction unambiguously;
2. should allow the untagged form to be reconstructed unambiguously;
3. should allow standard spellings to be used while preserving the spelling of the text; and
4. should be simple.

A simple system meeting these needs involves five special symbols, namely $\{ \} ^ { \wedge } \text{ ` } =$. Applied to standard transcriptions of classical Malay, it produces tagged text which is still comprehensible, though ugly, and which can easily be further edited if ever this should be required. Another simple program allows unambiguous reconstruction of the untagged text. The method is adequately illustrated in the following text excerpt from Mulyadi's edition of *Hikayat Indraputra* (1983):

1. Unlemmatized form

<P 49>

<L 1> <Q 1> Bahwa ini kisah ceritera Hikayat Indraputra yang indah-indah

<L 2> perkataannya yang masyhur pada tanah manusia dan pada tanah jin,

<L 3> terlalu elok rupanya, syahdan kesaktiannya, dan terlalu pantas

<L 4> barang lakunya, dan sikapnya terlalu baik dan rupanya terlalu amat

<L 5> manis seperti laut madu, dan jejaknya sederhana, barang lakunya

<L 6> dan pekertinya terlalu baik.

<L 7> Sebermula pada zaman itu seorang pun tiada samanya, dan terlalu

<L 8> arif dan bijaksana dengan gagah dan perkasyanya, dan <Q 2> dengan

<L 9> beraninya, dan beberapa pekerjaan yang tiada dapat dikerjakan orang

<L 10> dapat dikerjakan oleh Indraputra.

2. Lemmatized and normalized form

<P 49> <L 1> <Q 1> Bahwa={bahawa} {ini} {kisah} {ceritera} H{hikayat} Indraputra={inderaputera} {yang} {indah}-indah^r2` <L 2> `per{kata}an`nya` {yang} {masyhur} {pada} {tanah} manusia={manusia} {dan} {pada} {tanah} {jin}, <L 3> `ter{lalu} {elok} {rupa}nya`, syahdan={syahadan} `ke{sakti}an`nya`, {dan} `ter{lalu} {pantas} <L 4> {barang} {laku}nya`, {dan} {sikap}nya` `ter{lalu} {baik} {dan} {rupa}nya` `ter{lalu} {amat} <L 5> {manis} {seperti} {laut} {madu}, {dan} {jejak}nya` {sederhana}, {barang} {laku}nya` <L 6> {dan} {pekerti}nya` `ter{lalu} {baik}. <L 7> S`se`ber{mula} {pada} {zaman} {itu} `se{orang} {pun} {tiada} {sama}nya`, {dan} `ter{lalu} <L 8> {arif} {dan} {bijaksana} {dengan} {gagah} {dan} perkasyanya={perkasyanya}, {dan} <Q 2> {dengan} <L 9> {berani}nya`, {dan} `be`ber{apa} pekerjaan=`per{kerja}an` {yang} {tiada} {dapat} `di{kerja}kan` {orang} <L 10> {dapat} `di{kerja}kan` {oleh} Indraputra={inderaputera}.

3. Reconstituted form

Bahwa ini kisah ceritera Hikayat Indraputra yang indah-indah perkataannya yang masyhur pada tanah manusia dan pada tanah jin, terlalu elok

rupanya, syahdan kesaktiannya dan
etc.

This method of tagging permits all non-standard forms to be retained in conjunction with the lemmatized standard forms. For convenience, the standard spelling is taken to be that prescribed by the Malaysian Dewan Bahasa dan Pustaka (1982), or a spelling analogous to the prescribed forms.

Using this lemmatization program, a new and versatile form of the base text is created which will sustain concordance-making while also serving many other analytical purposes.

Forming the concordance

For the reason developed above, the conventional KWIC format is well suited to a general-purpose concordance of classical Malay. It is preferable that listings of identical key words should be sorted according to their context, rather than simply arranged in order of their occurrence in the text. Sorting by context, whether left or right, has the potential for grouping similar phrases and quasi-compounds, and uncovering formulaic features.¹⁶ Illustrated below is a right-sorted KWIC format resembling Packard's *Concordance of Livy* (1968), with some of the improvements suggested by Fleury (1986: 239). The text is again drawn from the first few pages of Mulyadi's *Hikayat Indraputra* (1983).

	barang	11	
	barang	dipeliharakan Allah taala juga tuhanku daripada bala raksaksa itu.	59.16
ada engkau bawakan aku obat beranak,	barang	kehendakmu kuberi. Jikalau aku beranak perempuan, kuberikan akan	57.1
berjalan. Ingat-ingat, jangan alpa	barang	kerja anakku.' Maka ujar Indraputra 'Ya ayahanda, sekalian hamba	57.34
laut madu, dan jejaknya sederhana,	barang	lakunya dan pekerjanya terlalu baik. Sebermula pada zaman itu	49.5
kesaktiannya, dan terlalu pantas	barang	lakunya, dan sikapnya terlalu baik dan rupanya terlalu amat manis	49.4
pun mungkin bertambah baik dan	barang	lakunya sedap manis. Wa 'llahu a'lam bi 's-sawab. Alkisah peri	53.50
-mongan berkeliling negeri, ujarnya,	'Barang	siapa kamu tahu obat dan hikmat beranak hendaklah kamu katakan,	55.3
inilah maka kubawa berjalan.' Maka	barang	siapa melihat rupa Indraputra itu kasih hatinya, ada yang memberi	52.44
peri orang kasih akan Indraputra itu,	barang	siapa memandang muka Indraputra itu tiada lupa daripada hatinya.	52.46
dua orang utas itu, 'Perbuatlah olehmu	barang	tahu kamu.' Maka diperbuat oleh pandai kayu itu seekor ikan terlalu	51.25
emas itu, 'Hai utas, perbuatlah olehmu	barang	tahumu supaya kulihat.' Maka pandai emas itu pun menyembah	51.33

Extending the principle of dictionary ordering, and also in order to make the display more effective, entries lying under a lemma-heading are best sorted into morphological groups. Thus efforts should be made to group together permutations of the verbal *di-/kau-/ku-* and *me-* forms (e.g. *dialami*, *mengalami*), and to keep secondary stems (e.g. *pengalaman*) and their derivatives (e.g. *berpengalaman*) together in the appropriate place within the main lemma grouping. This reflects normal dictionary practice. The listing order should also deal as well as possible with possible morphological ambiguities. For example, it is appropriate to list the *di-/kau-/ku-* verb forms directly after the unaffixed root because, with an explicit

¹⁶ Compare the following excerpt, which is sorted by right context, with the concordance excerpt given by Koster and Maier (1982:13), in which entries are sorted by order of appearance in the text.

subject, this verb form is that of the simple root. This departs from the alphabetic practice of the lexicographers, who would have *ber-* forms intervening.

With these purposes in mind, the key words within a lemma entry might be ordered as follows:

root
 di — kau — ku — me(N) —
 di — *i etc.*
 di — kan *etc.*
 ter — [stem group]
 ber — [stem group]
 per —
 diper — *etc.*
 diper — *i etc.*
 diper — kan *etc.*
 per — an [stem group]
 pe(N) — [stem group]
 pe(N) — an [stem group]
 ke — [stem group]
 ke — an [stem group]
 se — [stem group]
 —an [stem group]

Within each stem group, the same pattern of ordering applies. For forms in every group and subgroup a further subordination applies:

- | | | |
|--------|----------------------|---------|
| 1. —ku | 2. reduplicated root | 3. —lah |
| —mu | reduplicated stem | —kah |
| —nya | reciprocal | —tah |
| | — -me — | —pun |
| | — -ber — | |
| | ber — - —an | |

When the concordance is based on a fully lemmatized tagged text, it is a simple matter to convert the morphological analysis into a system for automatic ordering in morphological groups.

Other information

Along with a concordance, it is good practice to include some elementary statistical information about the text. Such information might include:

1. the type-token ratio, a simple calculation of the number of word-forms ('types') as a proportion of the total number of words ('tokens') in the text;
2. a frequency profile showing the number of types and tokens occurring at each level of frequency (3,000 types and tokens occurring only 1 time in the text, 1,000 types comprising 2,000 tokens occurring only 2 times in the text, etc.);

3. a list of types in order of frequency;
4. a list of types in alphabetical order giving frequency and rank;
5. for verse, a list of rhyme words sorted alphabetically by their terminations, with frequencies.

In conventional concordances, these statistics relate to *words* as they occur in the text. Assuming that the concordance of a Malay text is lemmatized, it would be both apt and valuable to generate such elementary statistics for morphemic units other than the word (Fleury 1986:242). So, all the above information in 1 to 4 could be supplied for roots and affixes as well as words. The frequencies of the affix types, in particular, might prove a telling discriminator of genre and antiquity.

The future

The particularities of one text become clearer when it can be put alongside others which have been treated in a comparable fashion. The value of one lemmatized text and one concordance is doubled by the creation of a second. In order to make comparisons more productive, and to enhance the contribution which computer-readable text can make to Malay philology, three areas of concern have to be borne in mind.

1. Questions relating to editorial practice and the state of Malay philology have been recently aired by Jones (1980) and Kratz (1981) and comprehensively discussed by Robson (1988:17-21). Jones and Kratz argue for conservative editorial practice. 'We have very far to go in our understanding of the early development of the Malay language, and our investigations can only be facilitated by editions of texts which produce clearly and accurately in easily comprehensible form the material of the manuscripts as it has survived.' An 'absolute minimum of alteration' is advisable (Jones 1980:126, 125). In the same vein, Kratz (1981: 238) notes the desirability of preserving 'all those peculiarities which may not seem of much significance within the limited framework of the particular text tradition, but which may well be important within a larger context'. In light of Behrend's experience, one such element may well be graphic form.¹⁷

Ideally, what we need now are diplomatic editions of single dated manuscripts of known provenance. Comparisons built upon such texts promise to add most to our understanding of historical linguistics, scribal cultures, and literary form.

2. The second concern relates to compatibility. Comparison is facilitated if texts have been treated in similar fashion. Two concordances arranged on the same principle are far more easily compared than two which follow different philosophies. An even more important concern is for the future use of the texts upon which current concordances may be based. If tagged

¹⁷ Consider the desirability of recording manuscript spelling forms; a system of diacritics may do so without disturbing the transliterated text, cf. Proudfoot 1967:i 16-17.

text from one source cannot readily be translated into the form of another source, the accumulation of banks of data is greatly hindered. Yet, if one cannot selectively tap material from diverse sources, comparative and historical work is needlessly handicapped. This is not a call for the straight-jacket. Compatibility does not require submission to any standard form, but rather an effort to maintain ready translatability between tagging systems. At the least, ambiguity must be avoided, and it must be possible to retrieve the untagged form of the text.

3. Thirdly, Kratz (1981: 240) notes the desirability of a 'generally accessible archive of computer-recorded texts'. As the computer is more and more used in preparing texts for publication, more computer-readable data will become available. An archive of the kind Kratz suggests, or at least a registry of resources, will help to avoid duplication of work, and to inform interested scholars of the availability of text materials. Fortunately such archives exist. One is the Oxford Text Archive, which acts as a depository for computer-readable text in a variety of European and other languages. It also acts as a registry or referral point for collections of computer-readable text held in other collections. Texts are made available for scholarly use on conditions determined by the depositors. The Oxford archive now holds substantial amounts of material in 35 languages. It has a negligible amount of Malay-Indonesian material: some extracts from Wilkinson and Winstedt's *Pantun Melayu* (1914), deposited by Thomas in conjunction with his investigation of prosody (1980, also 1979). There is no reason why the Oxford archive should not become a depository for further Malay material. In any case, to have some single registry or clearing-house will benefit future workers in this field.

In other fields of classical philology, the journey into computation has been under way for some little time. The University of California has assembled the *Thesaurus Linguae Graecae*, a collection of over 60 million words covering 'essentially all ancient Greek text materials extant from the period between Homer and AD 600', and this is now available on CD-ROM.¹⁸ It is presumptuous to think now of emulating this achievement in the field of classical Malay. But the example is there. Let us hope at least that future text editions will be prepared with the interests of computer analysis in mind. In today's world, publication of a text edition should mean not just the appearance of a book, but also of computer-readable disks or tapes.

¹⁸ See status report *Thesaurus Linguae Graecae* (1987). Even more ambitious, but not a project solely for classical literature, is the *Trésor général des langues et parlers français*, with 160 million words spanning the seventeenth to the twentieth centuries (Dendien 1988).

ABBREVIATIONS USED

BKf: *Bijdragen tot de Taal-, Land- en Volkenkunde* of the KITLV

KITLV: Koninklijk Instituut voor Taal-, Land- en Volkenkunde [Royal Institute of Linguistics and Anthropology], Leiden.

BIBLIOGRAPHY

- Akehurst, F.R.P., 1981, 'Concordances of troubadour poetry in Old Occitan', in: P. Patton and R.A. Holoien (eds), *Computing in the Humanities*, ch. 9, Aldershot: Gower.
- Becker, A.L., 1979, 'The figure a sentence makes: An interpretation of a classical Malay sentence', *Syntax and Semantics* 12:243-259.
- Behrend, T.E., 1987, *The Serat Jatiswara; Structure and change in a Javanese poem, 1600-1930*, doctoral dissertation, Australian National University.
- Benjamin, G., 1988, *Grammar and polity; The cultural and political background to standard Malay*, Singapore: National University of Singapore, Department of Sociology, Working Paper No. 88.
- Bessinger, J.B., Jr., with Philip H. Smith Jr., 1969, *A concordance to Beowulf*, Ithaca, N.Y.: Cornell University Press.
- Bessinger, J.B., Jr. (ed.), and Philip H. Smith Jr. (programmer), 1978, *A concordance to the Anglo-Saxon poetic records*, Ithaca, N.Y.: Cornell University Press.
- Brakel, L., 1975. *Hikayat Muhammad Hanafiyah*, The Hague: Nijhoff. [KITLV, Bibliotheca Indonesica 12.]
- Burton, Dolores M., 1981, 'Automated concordances and word-indexes: The process, the programs, and the products', *Computers and the Humanities* 15: 139-154.
- , 1982, 'Automated concordances and word-indexes: Machine decisions and editorial revisions', *Computers and the Humanities* 16:195-218.
- Cameron, A., Roberta Frank, and John Leyerle, 1970, *Computers and Old English concordances*, Ontario: University of Toronto Press.
- Dendien, Jacques, et al., 1988, 'Le Trésor général des langues et parlers français de l'Institut National de la Langue Française (I.Na.L.F.)', *Computers and the Humanities* 22:67-75.
- Dewan Bahasa dan Pustaka, 1982: *Daftar ejaan rumi Bahasa Malaysia*, Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Fleury, Ph., 1986, 'Du texte latin à la concordance imprimée', *Computers and the Humanities* 20:239-245.
- Fogel, Ephim G., 1962, 'Electronic computers and Elizabethan texts', *Studies in Bibliography* 15:15-31.
- Gärtner, Kurt, 1980, 'Concordances and indices to Middle High German', *Computers and the Humanities* 14:39-50.
- Hart, Thomas E., 1979, Review of Bessinger 1978, in *Computers and the Humanities* 13:229-244.
- Hockey, Susan, 1980, *A guide to computer applications in the humanities*, Baltimore: Johns Hopkins University Press.
- Howard-Hill, T.H., 1969-72, *Oxford Shakespeare concordances*, Oxford: Clarendon Press.
- , 1979, *Literary concordances*, New York: Praeger.
- Jones, Russell, 1980, 'Problems of editing Malay texts, discussed with reference to the Hikayat Muhammad Hanafiyah', *Archipel* 20:121-127.
- , 1983, *Hikayat Sultan Ibrahim; The short version of the Malay text*, Dordrecht: Foris. [KITLV, Bibliotheca Indonesica 24.]
- Josselin de Jong, P.E. de, 1961, 'Who's who in the Malay Annals', *Journal of the Malayan Branch of the Royal Asiatic Society* 34-2:1-89.
- Kassim Ahmad (ed.), 1964, *Hikayat Hang Tuah*, Kuala Lumpur: Dewan Bahasa dan Pustaka. [Siri Sastera DBP 17.]
- Klinkert, H.C., 1885, *Nieuw Maleisch-Nederlandsch woordenboek*, Leiden: Brill.

- Koster, G.L., and H. Maier, 1982, 'Variation within identity in the *Syair Ken Tambuhan* examined with the help of a computer-made concordance', *Indonesia Circle* 29:3-12.
- Kratz, E.U., 1981, 'The editing of Malay manuscripts and textual criticism', *BKI* 137:229-243.
- Lusignan, S., 1980, Review of Howard-Hill 1979, in *Computers and the Humanities* 14:129-130.
- Matheson, Virginia, 1979, 'Concepts of Malay ethos in indigenous Malay writings', *Journal of Southeast Asian Studies* 10:351-371.
- , in the press, *Tuhfat al-Nafis: A Malay Islamic history: The edited Malay text with introduction*, Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Matheson, Virginia, and B. Andaya, 1982, *The Precious Gift (Tuhfat al-Nafis)*, Kuala Lumpur: Oxford University Press.
- Micro-OCF*, 1987, Oxford Electronic Publishing, Oxford University Press. [The micro-computer version of *Oxford Concordance Program* version 2.]
- Mulyadi, S.W.R. (ed.), 1983, *Hikayat Indraputra; A Malay romance*, Dordrecht: Foris. [KITLV, Bibliotheca Indonesica 23.]
- Oakman, Robert, 1979, 'The Cocoa Concordance Package', Appendix, in: T.H. Howard-Hill 1979:77-80.
- Ong, Walter J., 1982, *Orality and literacy: The technologizing of the word*, London: Methuen.
- Packard, David W., 1968, *A concordance to Livy*, Cambridge, Mass.: Harvard University Press.
- Parrish, S.M., 1962, 'Problems in the making of computer concordances', *Studies in Bibliography* 15:1-14.
- Proudfoot, I., 1967, *The Hikayat Pelanduk Jenaka: Malay myth and oral composition*, Canberra: Faculty of Asian Studies, Australian National University (cyclostyled).
- Robson, S.O., 1988, *Principles of Indonesian philology*, Dordrecht: Foris. [KITLV, Working Papers 1.]
- Shellabear, W.G., 1915, 'Hikayat Seri Rama', *Journal of the Straits Branch of the Royal Asiatic Society* 71:1-285.
- Skinner, C., 1963, *Sja'ir perang Mengkasar*, The Hague: Nijhoff. [KITLV, Verhandelingen 40.]
- , 1965, *The civil war in Kelantan in 1839*, Kuala Lumpur: Malaysian Branch, Royal Asiatic Society. [Monographs 2.]
- , 1985, *The battle for Junk Ceylon; The Syair Sultan Maulana*, Dordrecht: Foris. [KITLV, Bibliotheca Indonesica 25.]
- Smith, Philip H., Jr., 1979, 'Programmer's preface', in: J.B. Bessinger (ed.), 1969:xxi-xxix.
- Spevack, M., 1973, 'Concordances: Old and new', *Computer Studies in the Humanities and Verbal Behaviour* 4:17-19.
- Sweeney, Amin, 1987, *A full hearing: Orality and literacy in the Malay world*, Berkeley: University of California Press.
- Tebben, Joseph R., 1977, *Homer-Konkordanz; A computer concordance to the Homeric hymns*, Hildesheim: Georg Olms.
- Teeuw, A., 1959, 'The history of the Malay language; A preliminary survey', *BKI* 115: 138-156.
- 'Thesaurus Linguae Graecae', 1987, *Literary and Linguistic Computing* 2:57. [Status report.]
- Thomas, Phillip, 1979, 'Syair and pantun prosody', *Indonesia* 27:51-63.
- , 1980, 'Long and short pantun lines', *Review of Indonesian and Malayan Affairs* 14:23-39.
- Tol, Roger G., 1984, 'Some order in ordering: Notes on *suruh*-constructions in 17th-century Malay', *BKI* 140:459-475.
- Tollenaere, F. de, 1976, 'Word indexes and word lists to the Gothic Bible: Experiences and problems', in: A. Jones and R.F. Churchhouse, *The computer in literary and linguistic studies*, Cardiff: University of Wales Press.
- Wilkinson, R.J., 1901-03, *A Malay-English dictionary*, 3 vols, Singapore: Kelly and Walsh.
- , 1932, *A Malay-English dictionary (romanised)*, 2 vols, Mytilene: Salavopoulos and Kindertlis.

Wilkinson, R.J., and R.O. Winstedt, 1914, *Pantun Melayu*, Singapore: Methodist Publishing House.

WordCruncher: WC Index & WC View, 1989 (version 4.30), Provo, Utah: Electronic Text Corporation, Brigham Young University.