CHAPTER 4
WORD CLOUDS AND POLITICAL SPEECHES

Zuraidah Md Don

4.1 INTRODUCTION

Word clouds have become increasingly familiar over the last two decades as a simple analogue representation of the content of a text. The frequency of items is indicated by size or colour, or a combination of both, and perhaps by positioning towards the centre or the periphery of a visual design. The cloud gives an immediate if vague indication of what the text is about. For example, if we see the item "Charles III" in large letters in the middle of a visual design and "republic" in smaller letters away from the centre, we can already begin to guess the content of the text on which it is based. Word clouds are associated with artificial intelligence (AI), and the details of any cloud depend in part on which of several available algorithms is used; but without knowing how the outcome is determined by the algorithm, it is impossible to get beyond a vague idea of the content of the text.

Word clouds have their counterpart in corpus linguistics in the form of word frequency lists, which have been routinely produced by corpus linguists over the last 50 years or so. Word lists have traditionally been created for beginning readers and language learners, and computer technology has enabled them to increase greatly in scale, the *magnum opus* being the Academic Word List (Coxhead, 2002), which is based on a corpus of about 3.5 million

words. Smaller projects undertaken in Asia have produced lists for specific disciplines, including food science and technology (Esfandiari & Moein, 2015), and social sciences (Chanasattru & Tangkiengsirisin, 2016); and also for different languages, for Thai (Ketmaneechairat & Maliyaem, 2023) and for Turkish (Aksan & Yaldır, 2012).

Corpus techniques have the potential to go far beyond the kind of information presented in word clouds and word frequency lists. The examination of clouds available on the internet shows that different algorithms process raw frequency (if at all) in different ways, which raises questions of validity and reliability. It is not at all clear that clouds measure what they intend to measure, or that different algorithms will generate similar clouds. Clouds are undoubtedly based on word frequency lists, but beyond that, it is not clear what theory if any they are based on. This lack of transparency reflects the status of AI algorithms as essentially "black boxes" so constructed that not even the designers will necessarily know how they work. In this case, linguistic processing offers not only transparency but also precision and clarity, because the processes involved have long been well understood, and different kinds of processing generate different frequency lists with known relationships to each other.

The extraction of information from word lists requires preliminary routine linguistic processing. The first stage is grammatical tagging, which involves associating a grammatical class (e.g., noun, verb etc.) with each word of the text, and the tags are used as input to syntactic analysis. For example, anyone who knows Malay also knows that *rumah* "house" tagged *kata nama* "noun" and followed by *besar* "big" tagged *kata sifat* "adjective" forms the phrase *rumah besar* "big house". This follows the grammatical rule that *kata sifat* follow *kata nama*, which is a very common rule of which huge numbers of instances can be found in any Malay corpus. Syntactic analysis opens up the possibility that phrases can be counted and included in frequency lists.

The linguistic analysis of texts necessarily starts with linguistic form, and the view taken for granted here is that the purpose of the close study of linguistic form is to gain access to the meaning. Syntactic analysis

provides the framework from which the meaning is to be inferred. Although much work in corpus linguistics is concerned with linguistic form, and in particular with the processes of grammatical tagging and syntactic analysis, linguistic analysis in general has long since gone on to study texts at the higher level of discourse. If corpus linguistics is to remain relevant for the long term, it has to cross the gap between the analysis of form and the analysis of discourse.

The initial investigation of the word list of the Mahathir Corpus (described below) indicated that it could indeed provide useful content information, and that it could not only reproduce the kind of analysis that lies behind word clouds, but also provide a much richer analysis based on linguistic theory, and a starting point for more ambitious linguistic analysis. The primary aim of this study is to test the informal hypothesis derived from this investigation. Although it is possible to make an intuitive judgement concerning what a text or corpus is about, it is essential for objective research to have something to measure which indicates the content. The simplest list ranks orthographic forms according to their frequency of occurrence, but a raw orthographic list does not give a useful indication of the content, and some processing is required of the linguistic form of the raw data. This paper identifies the kinds of linguistic processing required, and some of the problems to be encountered along the way.

One of the attractions of working on a single-authored corpus, as in the present case, is that it raises the possibility of creating a profile of the author corresponding to a word cloud. It is concerned with the speeches delivered by Tun Mahathir Mohamed before he stepped down as Prime Minister of Malaysia in October 2003 ("the Mahathir Corpus"), of which notwithstanding the contributions of speech writers, he can legitimately be regarded as the author. After two decades, not only is Tun Mahathir still at the time of writing a significant figure in Malaysian politics, but subjective evaluations of his role still range from one extreme to another. While his supporters might regard him as the man who created modern Malaysia or the best Prime Minister Malaysia has ever had, critical comments sometimes found in the local press also