The Pulse of the Text:
Using Digital Tools for Closer Reading

Ramon Guillermo

ABSTRACT

The use of computer-aided techniques in literary and cultural analysis has seen a surge in popularity in recent years due to the increasingly widespread availability of digitized textual and cultural materials. The Marxist literary scholar Franco Moretti (1950 -) is an example of this trend. However, the actual utility of many of these methods has not exactly been clear or obvious to students, scholars and researchers in the literary field. This paper will attempt to present and explain the intuitive utility of a set of digital tools which can be used in the analysis of textual materials. In contrast to the so-called "distant reading" advocated by Moretti which require the availabiity of massive digital corpora which are no yet available for many national literatures, these tools are examples of techniques which, in combination with traditional "close reading" can arguably lead to close(r) readings of texts. This paper will use Pramoedya Ananta Toer's (1925-2006) novel *Bumi Manusia* (1980) and its Filipino and English translations to serve as examples in translation analysis.

I. Introduction

This essay is an attempt to provide a step-by-step guide to the use of an experimental digital tool for textual analysis. A program written in the computer language Python (https://www.python.org/) will produce new-word, hapax legomena, repetition, lexical complexity and average betweenness centrality graphs for individual textual entities. (However, simplicity of exposition requires that only the first three types of graph be discussed below.) To give a more or less comprehensive idea of the possible range of applications of this digital tool, an original text in Bahasa Indonesia will be analyzed along with its English and Filipino language translations.

II. Graphing Texts

Modern lexicography has established that the frequency of co-occurrence or "collocation" of lexical elements in language use can serve as data for the analysis of word meaning. The empirically observable (or measurable) strength of connection between lexical elements in a text is known as "cohesion." However, the more conventional methods of tracking cohesion, e.g., concordance analysis, have usually been limited to the consideration of words in pairs. Gilbert Youmans (1991; 1994) demonstrated quite some time ago the utility of new-word graphs in the study of a type of lexical cohesion which goes beyond the limitations of traditional collocation analysis. In the intriguing approach he developed, a graph is produced by moving a "scan window" of variable length through a text one word at a time until it reaches the last word at the end of the text. The program counts the number of words in each successive scan window which appear for the first time in the text. All words which have previously occurred in the text will not be counted. Likewise, words which appear in a window multiple times will only be counted once. Since function words and other frequently occurring words aren't filtered out, the scan windows at the beginning will naturally produce values much higher than all other succeeding scan windows.

The theory is that valleys and peaks in such graphs can be used to detect thematic (or topic) boundaries and transitions in texts. The generation of the abovementioned new-word graphs provide interesting textual information because these make visible the phenomenon of "inter-

collocation" which goes beyond the pair-wise approach of conventional collocation analysis. Peaks in new word graphs indicate points where whole clusters of words, rather than just pairs, rise suddenly to the textual surface. According to Michael Stubbs,

> Youmans's work is a neat example of an all too rare event in linguistics: a previously unknown phenomenon which can be observed with a relatively simple technique. His method works, and provides replicable findings, in a small area of the social world which was previously thought to be closed to systematic study, it makes visible a kind of linguistic patterning which was previously invisible and unsuspected. (2002, 143)

III. Description of the Computer Programs

Two programs, "textanalysis-1.py" and "textanalysis-2.py" (See Appendix A), will be described below which generate data for new word, hapax legomena and repetition graphs. Data for two other types called "lexical complexity" and "average betweenness centrality" graphs will also be generated by the programs and will be described in passing below. Users of these programs should install the Python language distribution known as Anaconda so that the required mathematical modules can be accessed (in particular, Networkx) (www.continuum.io/anaconda). These two program files can be loaded or typed in and executed within the Spyder user interface included in the Anaconda package. The program files (with file extension ".py") must be placed in the same folder as the text data to be processed.

Important information for the execution of both programs should be typed in by the user at the blanks indicated at the beginning of the program code when these have been opened in the Spyder interface. The user must specify the file name of the ".txt" file to be processed (which should be in the same folder as the program to be executed) and the length of the scan window. The length of the scan window is still an experimental matter but mostly depends on the length of the text to be processed as well as the desired degree of resolution. A longer text will generally require a longer scan window. Overall, the setting of the length of the scan window

depends on obtaining a balance between the extremes of being swamped in detail or losing too much of it. As a rule of thumb, the user is encouraged to experiment for longer texts such as novels with scan window lengths of 250, 500, 1,000 and 2,000. In addition to the abovementioned information, "textanalysis-2.py" requires that a file called "stoplist.txt" be placed in the same folder as the text file to be analyzed. The "stoplist.txt" should contain a list of words which the user wished to filter out of the analysis. These are usually the frequently occurring function words of the language of the text to be analyzed. Function words and other words of low significance for the study among the one hundred most frequent words of the whole text could be included in the stoplist. The stoplist should just be a continuous listing of words separated by spaces with all words in lower case.

If the text(s) are not yet available in digital format, the encoding of the text(s) to be analyzed can be done manually or with available optical character recognition (OCR) software. However, with the latter option, it is still necessary to proofread the automatically encoded text since accuracy issues still arise given current limitations in the technology. It is suggested that two copies of the encoded text be maintained. The first version should be as close to the formatting of the original text as possible while the second is the same text which has been reformatted for the purposes of analysis. The reformatting should proceed as follows: (1) all punctuation and non-alphanumeric characters should as much as possible be deleted; (2) all tabs, carriage returns and multiple spaces between words should be removed; (3) if feasible, all characters should be reduced to lowercase (this may represent a problem for languages such as German); (4) files should be saved with a ".txt" extension. Some languages would be more amenable to these reformatting steps than others so it necessary to experiment in such cases.

The first output of the first program ("textanalysis-1.py") is a text file (file extension ".txt") with a name which includes the first three words of the text file under analysis and the setting of the length of the scan window (this will be the naming convention for all files generated by the two programs) (filename begins with "1-TEXT"). The beginning of the text file will include information on the total word count ("tokens"), the number of unique words in the text ("types") and the lexical complexity of the text (number of types divided by the total number of tokens).

However, the main content of the file is a list of words numbered and listed according to their order of appearance in the text. This file will be useful as a reference when looking up the contents of specific scan windows. The second output will be a text file (".txt") which is another numbered list of words listed according to their first appearances but with their subsequent reoccurrences blanked out (filename begins with "2-WORDFREQS"). The column to the right of the words indicates the frequency of appearance of each word which appears in the list. This file can be sorted in any spreadsheet program to obtain ordered frequency lists of the text being analyzed. The third output of the program will be a "comma-separated" spreadsheet file (with extension ".csv") with three columns (filename begins with "3-NEWWORDS"). The leftmost column contains the scan window number, the middle column contains the total number of new words counted within that scan window, and the rightmost column contains the number of new words in the window expressed as a fraction of the length of the scan window (total new words divided by the length of scan window). The range of values in the second column will be from zero to equal the length of the scan window while the range of values in the third column will be between zero and one. Obviously, the last scan window number in the leftmost column will be equal to the total length of the text minus the length of the scan window. The fourth output of the program will be another spreadsheet file (".csv") which has three columns (filename begins with "4-HAPAXA"). The leftmost column contains the scan window number, the middle column contains the total number of single occurrence words (or "hapax legomena") counted per scan window, and the rightmost column contains the number of hapax legomena in the window expressed as a fraction of the length of the scan window.

The first output of the second program ("textanalysis-2.py") is a spreadsheet file (".csv") containing textual "repetition" data (filename begins with "5-REPETITION"). In the same manner as the first program, this is produced by passing a scan window of variable size one word at a time through a text from beginning to end. The program totals the frequency of occurrence in the whole text of each word per scan window. This type of repetition data can show the areas in the text where the most "popular" words in the text as a whole occur. High points in the repetition graph may therefore be read as textual zones with a high cohesive relation to the text taken as a whole rather than with any of its specific parts. Obviously,

areas of the text where many function words occur together will have high values. This is mitigated by a stoplist function which screens out specified function words and other nonsignificant frequently repeating words (listed in the "stoplist.txt" file) from the count. The format of the repetition data has three columns. The first column contains the scan window number, the second contains the aggregate number of frequencies of occurrence of all words in the scan window (excluding words in the stoplist), while the third column contains the value in the second column divided by the total number of scan windows and therefore stands for the average rate of appearance of the words per scan window in the text as a whole.

The second output of the "textanalysis-2.py" program is a file (".csv") contains lexical complexity data (filename begins with "6-LEXCOMP"). It was mentioned above that the first program ("textanalysis-2.py") computes lexical complexity for the text as whole by dividing the number of types by the total number of tokens. The range of values for lexical complexity is from "most complex" (1.0), wherein the total number of unique words (types) is equal to the total number of words (tokens), to least complex (approaching zero). Calculating lexical complexity as the scan window of variable length moves along the text can more specifically indicate the areas of the text where lexical complexity is high or low. Areas of high lexical complexity indicate points in the text where the vocabulary is particularly rich and may therefore be useful for some types of textual analysis. Areas in the text with a high occurrence of new words or hapax legomena may also be areas of highly complex vocabulary. The first column in the spreadsheet file contains the scan window number while the second contains the value for lexical complexity.

The third output is a spreadsheet file (".csv") containing the "average betweenness centrality" value for all types within the span of the moving scan window (filename begins with "7-BETCEN"). When the text as a whole is taken as a complex network of interconnected words or "nodes," "betweenness centrality" is a measure of the degree or strength by which a "node" serves to connect other nodes and clusters of nodes to each other. The output spreadsheet file consists of two columns wherein the first is the scan window number while the second indicates the average betweenness centrality values of all the nodes within the scan window. To serve as a reference for the analysis of the betweenness centrality graph, a fourth file

(".csv") will be outputted which contains three columns (filename begins with "8-BETCENLIST"). The first is the scan window number, the second will be the words in the file in their order of appearance in the text, while the third will be the betweenness centrality value corresponding to each word listed in sequence.

IV. Step-by-step Translation Analysis: Pramoedya Ananta Toer's *Bumi Manusia* (BM)

Although the examples below will be from translation analysis, applications to other areas of literary and textual analysis should be immediately evident to the reader (the first application to translation studies is Guillermo (2009b)). In spite of their general applicability, the graphs seem to be particularly useful for translation analysis since these can serve as language-independent bases of comparison between the source and target texts in translation. The tools are "language-independent" because they do not discriminate in any way between languages in producing their output. The tools also do not bring any presuppositions about grammar, morphology or semantics to the texts to be analyzed.

The great Indonesian novelist Pramoedya Ananta Toer's (1925-2006) (1980) *Bumi Manusia* (BM) and its English and Filipino language translations will serve as the main examples in the analysis below. BM is the first volume of Pramoedya's famous *Buru Quartet* which is said to have been originally composed orally while Pramoedya was imprisoned by the dictator Soeharto on Buru Island from 1969 to 1979. A runaway bestseller in its first printing, it was banned in Indonesia soon after its year of first publication in 1981. UNESCO's Index Translationum (http://www.unesco.org/xtrans/) lists various translations of this novel into English, German, Japanese, Korean, Dutch, Spanish, Norwegian and Swedish from 1982 to 2004. More pertinent for the present study are Max Lane's (1996) English translation, *This Earth of Mankind* (TEOM), which was first published in 1982 by Penguin Books, and Thelma Kintanar's (1989) Filipino translation entitled, *Ang Daigdig ng Tao* (ADNT), which was printed by the Solidaridad Publishing House in 1989.

BM, TEOM and ADNT were digitized using a combination of OCR software and manual encoding. Two versions were maintained for each text. One version was meant to stay as close as possible to the original

formatting of the texts (in ".doc" or ".rtf" formats) while the other version was processed to facilitate automated analysis by removing or deleting all formatting such as the following: punctuation marks, non-alphanumeric characters, double spaces, tabs, carriage returns and reducing the whole text to lowercase. The latter versions were saved in ".txt" format using the initials of the titles as filenames "bm", "teom", and "adnt." These files were then placed in individual folders together with the programs textanalysis-1.py and textanalysis-2.py. After the programs were run, the files generated in ".csv" format labeled 3-NEWWORDS, 4-HAPAXA and 5-REPETITION were opened in a spreadsheet program for visualization and analysis.

A. Variable Length Scan Windows for BM

Figure 1 shows new word graphs for BM with different scan window lengths of 250, 500, 1000 and 2000 (by plotting the output file 3-NEWWORDS, set at four different scan lengths). It can be observed that although the different scan lengths exhibit the same general form for a single text, the features become more pronounced as the scan window gets longer. The degree of resolution can therefore be adjusted to match the desired level of detail by experimentally varying the length of the scan window. For the purposes of this discussion, the level of detail provided by setting the scan window length at 500 has been considered sufficient.

The occurrence of high values at the starting point of new-word graphs is to be expected since all types are counted as new words at the beginning of the scan process. For example, the 500-word scan window for BM from the first word to the five hundredth registers 311 new words. This means that unique types make up 62.2% of the words in the first scan window while the remainder consists of repetitions of these newly introduced words which may subsequently resurface at different rates.

B. Combined New-Word, Hapax Legomena and Repetition Graphs for BM

Figure 2 shows the combined new-word (blue), hapax legomena (red) and repetition (green) graphs for BM with a 500-word scan window setting for all three. (The plotted values on the new-word and hapax legomena graphs are the wordcounts per scan window divided by the scan window length, or in this case, by 500.) The visible relationship of

new-word occurrences to hapax legomena occurrences can be observed in this graph. Obviously, the hapax legomena graph cannot have a greater value than the new-word graph at any point since single occurrence words practically make up a subset of the total number of new words for each scan window. It can also be seen that the hapax legomena graph follows the peaks and valleys of the new-word graph for the most part even though the percentages of hapax legomena occurring in each window can vary quite significantly. The general trend seems to be that the hapax legomena and the new words graphs begin to overlap almost completely as one moves towards the latter third of a text since most new words at this point will no longer have the occasion to resurface in the succeeding parts of the text. In general, hapax legomena graphs can help detect lexical islands in the text which pertain to themes or topics which arise only once.

On the other hand, intercollocated lexical elements which continue to repeat may exert a further generative effect on the text as they resurface in various subsequent parts of the text. To aid in the investigation of the latter phenomena, a "repetition graph" which is also produced by scanning the text from beginning to end can show the points in the text which contain the lexical items which repeat most frequently throughout the text as a whole (excluding the most frequently occurring function words). While the values for the new-word and hapax legomena graphs range between 0 and 1, the value for the repetition graph is the average number of appearances of the lexical items in each respective scan window in all scan windows (the total number of repetitions of all the lexical items in the scan window divided by the total number of scan windows). Other more refined techniques for detecting repetition phenomena can be devised. For example, generating the average lexical overlap of each scan window with all other scan windows in the text seems to present a good alternative (Hoey 1991). However, this latter method requires much more processing power than is generally available especially for longer texts such as novels. The relationship between the new-word graph and the repetition graph is still a matter of study, but these do not at the outset, exhibit clear patterns of correspondence. It seems to be the case, however, that there is a roughly inverse relationship between hapax legomena graphs and repetition graphs. This is intuitively plausible since these two graphs exhibit the opposed logics of non-repetition and high repetition. The final high peak at the end of the BM repetition graph (point 94593 with a value of 0.297) indicates

the probable recapitulatory nature of the last sections of the text (this was also remarked upon in Guillermo (2009b)).

C. Combined BM, TEOM and ADNT Graphs

Figure 3 shows the combined new-word graphs of BM, TEOM and ADNT with the 500-word scan window settings (the x-axis indicates values for BM only). Since both TEOM and ADNT are longer texts than BM (see Table 1), both had to be rescaled, resulting in a certain degree of distortion. Nevertheless, rough and somewhat irregular correspondences between the peaks and valleys of TEOM, BM and ADNT can be observed throughout the graph. The scatter-plots of BM with TEOM and ADNT shown in Figure 4 are useful for visualizing their respective degrees of correspondence. The tendency of the scatterplots to form a straight line indicates degree of correspondence which cannot be purely random. One infers from this that new-word graphs exhibit a degree of stability across translations (Cabatbat et al. 2014). (Tests for Pearson's correlation shows high statistical significance with p < .001 for all pairs BM and TEOM, BM and ADNT and TEOM and ADNT.) The combined hapax legomena and repetition graphs of BM, TEOM and ADNT shown in Figures 5 and 6, exhibit some correspondences which seem to be irregular and contingent as opposed to the relative regularity observed in the new-word graphs (scatterplots also do not reveal relationships of correlation). An interesting feature is the valley at point 60204 of BM where no new words are introduced (y-axis value is 0). A corresponding dip does not occur in TEOM and ADNT.

Now that the general relationships have been sketched out, it is possible to zoom in to particular points or sections of the text.

D. New-Word Peaks and Text Alignment for BM and TEOM

Three peaks in the BM new-word graph can serve as examples for demonstrating the utility of the graphs which have been generated. These are points A, B and C corresponding to BM scan windows numbered 18975 (with 99 new words), 29333 (with 81 new words) and 34602 (84 new words). Figure 7 indicates the location of these points on the BM new-word graph with the hapax legomena and repetition graphs included for comparison. All three of these areas on the BM graph demonstrate the

occasionally inverse relationship between hapax legomena and repetition graphs. For example, at approximately BM point 20907, the hapax legomena graph peaks while the repetition graph dips drastically. Figure 8 shows the corresponding points in BM (18975, 29333, 34602) and TEOM (23177; 36438; 43657).

The extraction of the contents of the 500-word scan window for all three corresponding points in BM and TEOM (using output files 1-TEXT and 2-WORDFREQS) show significant overlaps in all three cases (see Table 2). Using BM as reference, points A for BM and TEOM start at almost exactly the same part of the text with the sentence in BM, "dia lahir pada hari pasaran *Paing*" and its direct translation in TEOM as, "he was born on the market day of *Paing*." Point B is just off by a few words with BM starting just just one sentence before the phrase "terlalu rendah untuk dibicarakan" (unworthy of discussion) and TEOM starting exactly on the words, "as worthy of discussion." The case of Point C is a bit more complex. TEOM starts with the phrase, "had once dreamed," while in BM, the same phrase "pernah aku impikan" (I have dreamed), occurs much later, 150 words into the middle of the scan window. These examples indicate that, although new-word graphs may be used for aligning two texts in a translational relationship with each other, its degree of accuracy for specific areas of text may also be highly variable.

E. The Analysis of New-Word Peaks: Seed-words and Major Themes

The extracted words which make up the new-word peaks exhibit certain interesting traits. It can be observed from the current example that new words with a frequency of one (hapax legomena) usually make up the majority of new words in each new-word peak as soon as the initial burst of new words at the beginning of the text has subsided. The present hypothesis is that the new words which have the subsequent highest rates of occurrence while moving forward in the text as a whole seem to be the ones which perform a strongly generative function in determining the dominant thematics of the new-word peak. As a matter of convention, and purely for heuristic purposes, the five words with the highest frequency from each new-word scan window will be extracted and labelled "primary seed-words." However, in the interest of simplification, only the nouns

among these most frequent words will be considered as properly "seed-words." (see Table 3.)

The primary seed-words which have been identified for BM are, "Sastrotomo" (the name of Sanikem's father; with 1 occurrence in the scan window itself and 14 appearances subsequently in the whole text), "juru-tulis" ("clerk" or, more archaically, "scribe"; 4 occurrences in the scan window and 15 appearances subsequently), "Sanikem" (the original name of the central female character of BM; 1 occurrence in the scan window and 16 appearances subsequently), and "ayahku" ("my father"; 1 occurrence in the scan window and 16 appearances subsequently). On the other hand, the seed-words for TEOM are as follows, "Tulangan" (the location of a sugar plantation on Java; 2 occurrences in the scan window and 10 appearances subsequently), "clerk" (3 occurrences in the scan window and 11 appearances subsequently), "Sastrotomo" (1 occurrence in the scan window and 14 appearances subsequently in the whole text), "relations" (1 occurrence in the scan window and 14 appearances subsequently in the whole text), and "Sanikem" (1 occurrence in the scan window and 18 appearances subsequently in the whole text).

Three seed-words appear in two connected sentence near the beginning of BM point A, "*Ayahku* bernama *Sastrotomo* setelah kawin. Kata para tetangga, nama itu berarti: *jurutulis* yang utama" (My father was named Sastrotomo after marriage. The neighbours said, this name means: the foremost scribe/clerk). This conjunction between "Ayahku" ("my father," named "Sastrotomo") and "jurutulis" ("clerk" or "scribe"), "triggers" the way whole section of text corresponding to point A converges within itself the thematic zones of family, community and honor, on the one hand, with those of employment and modern production relations on the other (Figure 9). In TEOM however, only one seed-word in the same sentences appears, "My father changed his name to *Sastrotomo* after he was married. The neighbors used to say the name meant the foremost scribe." However the seed-words "Sastrotomo" and "clerk" (as with "relations" and "Sanikem") arguably trigger the same processes of convergence of lexical thematics (Figure 10). The striking conjunction between the discourses of individual and family "honor" and "promotion" (in employment) occurs in a sentence in BM, "Ia impikan *jabatan* lebih tinggi sekali pun jabatan-nya sudah cukup tinggi dan *terhormat*" (He dreamed of a higher position/

post even though his current position was high enough and respected/honored). Figure 11, shows just how rich this particular section is with respect to this convergence of themes in BM. The same conjunction occurs in TEOM, "He dreamed of a higher *post* even though the job he held was quite a *respected* one." This logic of equating promotion with a rise in social status is even more explicit in the phrase, "kenaikan jabatan, kehormatan dan ketakziman" (promotion in position, honor and status), wherein a "rise" (kenaikan) in one's employment occurs simultaneously and in parallel with a "rise" in honor and status. This is translated in TEOM as, "rise in position, respect or esteem." BM point A includes words such as "upah" (salary), "gaji" (wage) and "uang" (money) which closely intertwine with words such as "hormat" (honor; respect) "takzim" (honor; respect, esteem) and "martabat" (status; dignity; prestige). These latter cluster of words are mirrored in Lane's translation by the triad of "respect," "esteem" and "dignity" (though the appearance of the word "dignity" already occurs beyond the 500 word boundary of TEOM point A).

F. Translational Semantic Shift: Honor to Respect

The main difference which arises between the relevant sections from BM and TEOM is that a certain shift seems to have occurred which can be articulated in English as a shift from the idiom of "honor" to "respect" (Table 4). It is true that these two words necessarily overlap in meaning and usage in the English language. For example, a popular definition of "honor" is "great respect." Nevertheless, it would be a mistake to conflate these two terms (Olsthoorn 2015). One difference is that "respect," as well as the concept of "dignity," can be generalized to encompass all humans such that we are called upon to uphold "human dignity" and exercise a properly egalitarian "respect for others." This evidently does not hold true for "honor" which is essentially an exclusive concept. If everyone were honored in the same way, the very notion would lose its meaning. Secondly, although "esteem" is arguably similar to "honor" in possessing a certain exclusivity in its application, the former term shares with "respect" to the characteristic of being dependent on recognition, whether by one's self, as in "self-respect" or "self-esteem," or by others, while "honor" has an element which is held to be irreducibly intrinsic to its possessor, who on this basis, expects it from others. Because of this nuance, "honor" lends itself more easily to concepts of ascribed status whereas "respect" dwells

more comfortably in the world of achieved status and social mobility. Thirdly, "honor," like "shame," has connotations which are more strongly collective, in the familial and communal sense, than "respect" with its more individualistic usages. Overall, it is said that in the English language the idiom of "respect" has taken over the formerly dominant, and in some respects already archaic, idiom of "honor." For example, the term "honor killings" could not be written as "respect killings" or even as "dignity killings."

On the broadest level of usage, the words "hormat" (or "kehormatan") and "takzim" (or "ketakziman") spans the range of meanings covered by the English words "respect," "esteem" and "honor." On the other hand, "martabat" (or "kemartaban"), borrowed from Arabic, covers the gamut of terms "rank," "status," "prestige," and "dignity" (Cf. Tan 1981; Saber et al. 1980; Riemer 1987). "Hormat" cannot simply be reduced to either "respect" or "honor" outside of its context of usage, and this context is quite often simply too complex to permit an easy disambiguation (Echols and Shadily 2016; Departemen Pendidikan Nasional 2016). The use of the triad "respect," "esteem" and "dignity" in the English translation results in the foregrounding of a more contemporary English idiom behind which has been occluded the idiom of "honor." The consequence is that the ironic humour which arises when Pramoedya pits traditional "family honor" against the shameless desire for "job promotion" and money in the context of modern economic relations loses much of its impact in the translation. As if to heighten the bitter humor of the situation, Pramoedya also related how Sastrotomo made use of prayer, fasting, magic and the mystical arts to attain his ambition of promotion to paymaster.

Given its already more strongly egalitarian and individualistic connotations, the English idiom of "respect" cannot so strongly bind the disparate discursive fields of family, community and honor, on the one hand, with those of employment, commodity economy and modern production relations on the other. Pramoedya's use of the idioms of "hormat," "ketakziman," and "martabat," on the other hand, more fully dramatizes the collision between the lived textures of tradition and the inroads of economic modernity. In fact, the whole novel ends with the word "hormat" ("Kita telah melawan, Nak, Nyo, sebaikbaiknya, sehormat hormatnya." Translated by Lane as, "We fought back child, Nyo, as well and honorably as possible."), which shows how BM point A perhaps

unexpectedly captures in miniature what may be the main dilemma of the novel as a whole. It should be emphasized that, more important than the apparent contradictory dualisms at play in these discursive phenomena, one should observe closely how these elements are closely imbricated and articulated with each other.

V. Some Reflections on the Problem of Translatability

In his classic essay on the Javanese concept of "power," Benedict Anderson included a strongly worded cautionary footnote,

> In the ensuing discussion of Javanese political ideas, I am attempting to map out a pure model for analytical purposes. Traditional Javanese political culture was an extremely complex phenomenon, in which, as in any other culture, it would be naïve to try to discern complete consistency. In that traditional culture an indigenous matrix was imperfectly compounded with heterogeneous Brahmanic, Buddhist, and Islamic elements. Nonetheless, the slow process of absorption and synthesis over the centuries prior to the "coming of the West" permitted the crystallization of a relatively high degree of internal consistency. The model I am trying to delineate is thus an "ideal type" which should not be taken as a historical reality… Java's subjection to Western political, economic and cultural domination has, particularly in the past hundred years, set in motion an irremediable process of decrystallization. Contemporary Javanese political culture is therefore a heterogeneous, disjunctive, and internally contradictory complex of traditional and Western elements, with a lower degree of internal logic and coherence than in the past… (1990, 20)

Despite the manifold processes of "decrystallization" which have already taken place in Javanese political culture, Anderson continued to make a case for incommensurability when it came to certain concepts, "When I say that the Javanese have a radically different idea of power from that

which obtains in the contemporary West, properly speaking this statement is meaningless, since the Javanese have no equivalent word or concept" (4). Aside from the problem of translating "power," one surmises that it was precisely the "high degree of internal consistency" and coherence of worldview which enveloped such Javanese concepts as "rasa," "budi," "nyawa" and "sampoerna" which made Anderson despair of finding any translational equivalents for these culture concepts in the existing idioms of English when he was himself translating a work by Pramoedya and other Indonesian writers (Guillermo 2017). Taking a cue from Anderson's insights, one could argue that Pramoedya, in the short text analyzed above, acutely dramatizes a process of "decrystallization" or destructuration of the originally tightly integrated conceptual cluster "hormat," "takzim" and "martabat" in the face of capitalist modernity. Moreover, the translational "flattening" of "hormat" as the equivalent of "respect" in English facilitates and smoothens the immediate process of translation while also participating in what may be conceived of as a bidirectional process of decrystallization.

The abovementioned considerations are the reasons why discussions of "untranslatability" which merely evolve around individual words and their etymologies isolated from their associated semantic fields (undergoing processes of decrystallization and crystallization) are inadequate and inconclusive (Guillermo 2009a; Guillermo 2016). On the other hand, an overly rigid privileging of internal consistency, logic and coherence of conceptual assemblages conceived as permanent impediments to translation, cannot fully take into account the dual historical dynamic of crystallization and decrystallization which in certain crucial ways respectively enable and disable translational processes.

VI. Conclusion and Recommendations

The present study has attempted to provide a step-by-step introduction to using a digital tool for the analysis of textual entities. Since almost all aspects of the method described above are experimental, further testing would be necessary to understand and explore their full potential and limitations. These techniques are at present conceived as being complementary to the more familiar qualitative approaches in close reading, textual analysis and social criticism. It could be the case that future rigorous critical and philological studies in the analysis of individual texts

might include these methods as a matter of course. What is more doubtful is that these methods in all their simplicity can become stand-alone and self-sufficient. The great Filipino literary scholar, Resil Mojares (2017, 110), expressed enthusiasm about the potential of these approaches in generating "fresh scholarship" in the humanities.

In the aggregate, all natural language texts fall in a range between least and most complex. However, within the texts themselves, there is a constant oscillation between the repetition of the old and the introduction of the new. This unceasing oscillation can be called, metaphorically, the "pulse" of the text. Norman Fairclough gives a useful guide in deepening an understanding of this phenomenon as being deeply embedded in the social nature of language,

> It is important to avoid a one-sided emphasis on either repetitive or creative properties of texts. Any text is part repetition, part creation, and texts are sites of tension between centripetal and centrifugal pressures. Texts will vary in the relative weight of these pressures depending upon social conditions so that some texts will be relatively normative whereas others are relatively creative. Centripetal pressures follow from the need in producing a text to draw upon given conventions, of two main classes; a language, and an order of discourse – that is, a historically particular structuring of discursive (text-producing) practices… Centrifugal pressures come from the specificity of particular situations of text-production, the fact that situations do not endlessly repeat one another, but are, on the contrary, endlessly novel and problematic in new ways… The tension between repetition and creation, centripetal and centrifugal pressures, manifests itself in varying degrees of homogeneity or heterogeneity of textual forms and meanings… The heterogeneities of texts code social contradictions. It is this property of texts that makes them the sensitive indicators of sociocultural processes and change I referred to above in discussing texture. Social contradictions may even be condensed into particular collocations in texts, particular patterns of co-occurrence

and mutual predictability between words, for instance, the collocation *enterprise culture*. The homogeneities/ heterogeneities of texts can be shown through intertextual analysis of the links between a text and other texts and text types, which is a necessary complement to linguistic analysis within the analysis of texts. (1995, 7-8)

The type of computer-aided "close(r)" reading being advocated here is also more feasible with respect to languages such as Filipino/Tagalog, Indonesian, Malay etc. which do not yet have large text copora comparable to the 4.5 billion word Bank of English (https://collins.co.uk/page/ The+Collins+Corpus), the 4 billion word German Reference Corpus (http://www1.ids-mannheim.de/kl/projekte/korpora.html) or even the 105 million word Kotonoha Japanese Language Corpus (http://www. kotonoha.gr.jp/shonagon/). Given the current limitations, some further research projects or applications might include the following:

- Refining the interpretation of both peaks and valleys in new-word graphs;
- Elaborating on the relationships between the different types of graphs produced by the programs;
- Tests for translational correlations for large numbers of translations in several languages of a single text;
- Tests for translational correlations for large numbers of translations into a single language of a single text;
- Testing for correlations between different texts of the same genre such as romance novels of roughly the same length and by the same publisher;
- Analysis of different types of texts, e.g., expressive or informative;
- Textbooks in different scientific disciplines can be used to test the ability of the tools to detect thematic boundaries
- Collections of texts of a single author arranged chronologically can be used to detect shifts in vocabulary usage which may be useful in mapping intellectual history.

Bibliography

Cabatbat, Josephine Jill T., Jica P. Monsanto, and Giovanni A. Tapang. 2014. Preserved Network Metrics Across Translated Texts. *Int. J. Mod. Phys.* C 25, 1350092.

Departemen Pendidikan Nasional. 2016. *Kamus Besar Bahasa Indonesia. Pusat Bahasa.* Jakarta: PT Gramedia.

Echols, John M., and Hassan Shadily. 2016. *Kamus Indonesia Inggris. Edisi ketiga yang diperbarui.* Jakarta: PT Gramedia.

Fairclough, Norman. 1995. *Critical Discourse Analysis: The Critical Study of Language.* Essex: Longman.

Guillermo, Ramon G. 2009a. Pantayong Pananaw and the History of Philippine Political Concepts. *Kritika Kultura 13*: 107-116.

_____. 2009bg. *Translation and Revolution: A Study of Jose Rizal's Guillermo Tell.* Quezon City: Ateneo de Manila University Press.

_____. 2016. Revolusi! Rebolusyon!: A Filipino Revisiting of Benedict Anderson's 'The Languages of Indonesian Politics' (1966). Southeast Asia Research Center (SEARC), City University of Hong Kong, Hong Kong, Working Paper Series no. 131.

_____. 2017. 'He was a translator': Benedict Anderson, Translation and Cosmopolitanism. In *Proceedings of the International Conference: On Reviving Benedict Anderson. Imagined (Cosmopolitan) Communities* Ed. by Budi Susanto, S.J. A. Windarto, A. Harimurti. Yogyakarta: Sanata Dharma University Press, 243-251.

Hoey, Michael. 1991. *Patterns of Lexis in Text.* Oxford: University Press.

Mojares, Resil B. 2017. *Interrogations in Philippine Cultural History*. Quezon City: Ateneo de Manila University Press.

Olsthoorn, Peter. 2015. *Honor in Political and Moral Philosophy*.New York: SUNY Press, 2015.

Pramoedya Ananta Toer. 1980. *Bumi manusia : sebuah roman*. Jakarta: Hasta Mitra.

_____. 1989. *Ang Daigdig ng Tao,* trans. Thelma Kintanar. Manila: Solidaridad Publishing House.

_____. 1996. *This Earth of Mankind,* trans. Max Lane. New York: Penguin Books.

Riemer, Carlton L. 1987. Maranao maratabat and the concepts of pride, honor, and self-esteem. *Dansalan Quarterly* 8(4): 125–80.

Saber, Mamitua, Mauyag M. Tamano and Charles K. Warriner. 1980. The Maratabat of the Maranaw. *Philippine Sociological Review* 8, no. 1/2: 10-15.

Stubbs, Michael. 2002. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell Publishing.

Tan, Armando L. 1981. Shame, Reciprocity, and Revenge: Some Reflections on the Ideological Basis of Tausug Conflict. *Philippine Quarterly of Culture and Society* 9, no. 4: 294-300.

Youmans, Gilbert. 1991. A New Tool for Discourse Analysis: The Vocabulary-Management Profile. *Language* 67, no. 4:763–89.

_____. 1994. The Vocabulary-Management-Profile: Two Stories by William Faulkner. *Empirical Studies of the Arts 12*, no. 2:113–30.

Table 1: *Bumi Manusia* (BM), *This Earth of Mankind* (TEOM) and *Ang Daigdig ng Tao* (ADNT)

Title and Language	Tokens	Types	Lexical Complexity (Types/Tokens)
BM (Bahasa Indonesia)	95251	10239	0.10749493443638387
TEOM (English)	124747	7961	0.0638171659438704
ADNT (Filipino)	126041	10845	0.08604343031235867
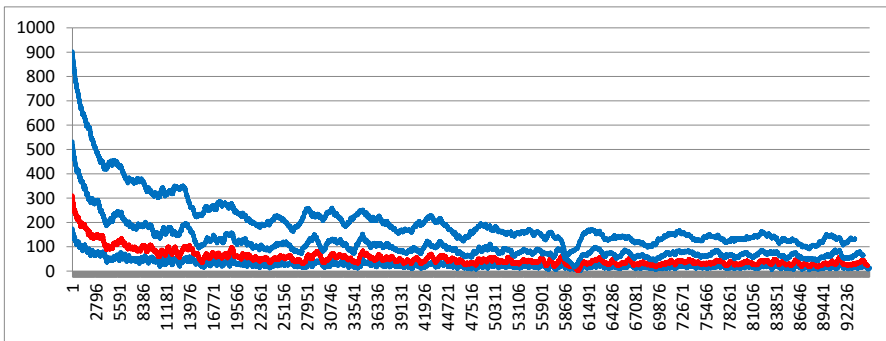
Figure 1: New-word graphs with different scan window lengths for BM from top to bottom: 2,000; 1,000; 500 (red); 250
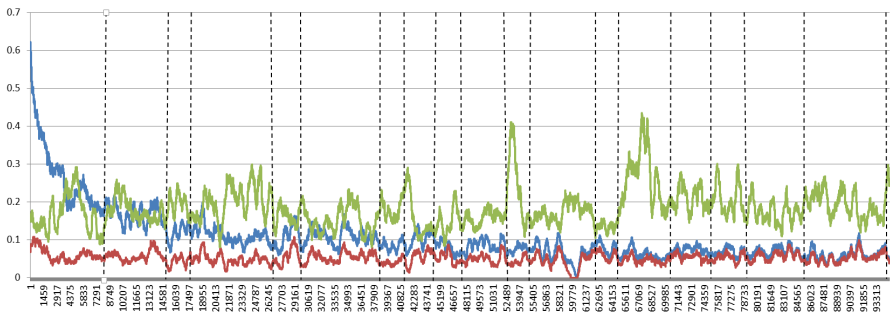
Figure 2: New-word graph (blue), hapax legomena graph (red) and repetition graph (green) for **BM** set at scan window lengths of 500 with values divided by scan window length (500) (dashed lines mark the start and end of chapters)
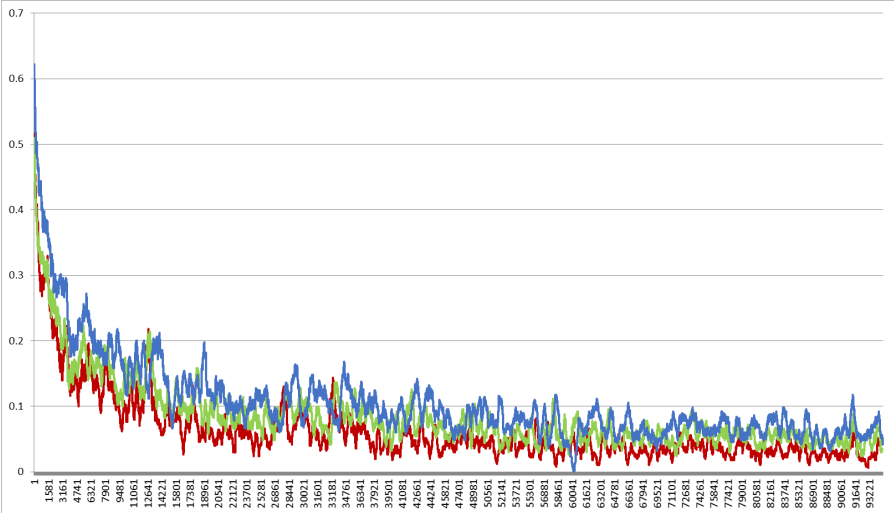
Figure 3: New-word graphs for BM (blue); TEOM (red) and ADNT (green) set at scan window lengths of 500 with values divided by scan window length (500)
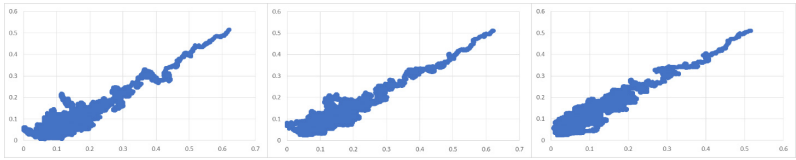
Figure 4: Scatterplots for BM and TEOM (left); BM and ADNT (middle); and TEOM and ADNT (right) (Pearson's correlation tests for all pairs exhibit high statistical significance with p < .001).
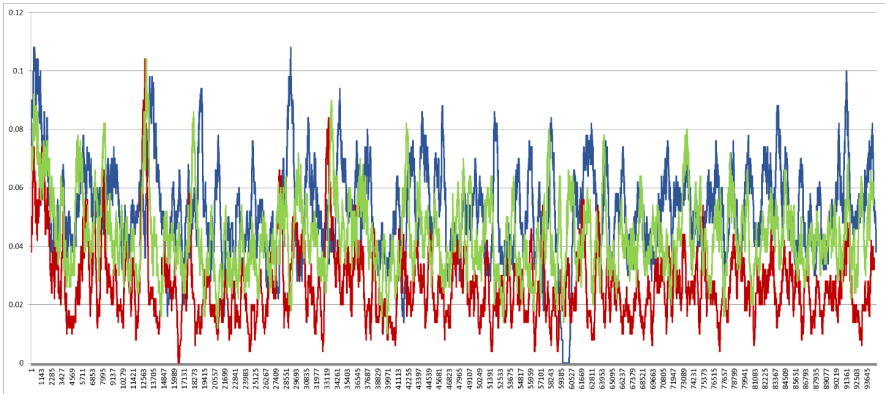
Figure 5: Hapax legomena graphs for BM (blue), TEOM (red) and ADNT (green) set at scan window lengths of 500 with values divided by scan window length (500)
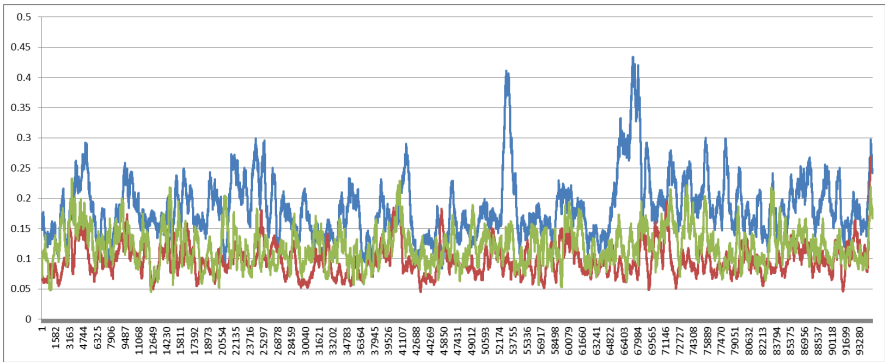
Figure 6: Repetition graphs for BM (blue), TEOM (red) and ADNT (green) set at scan window lengths of 500 with values divided by scan window length (500).
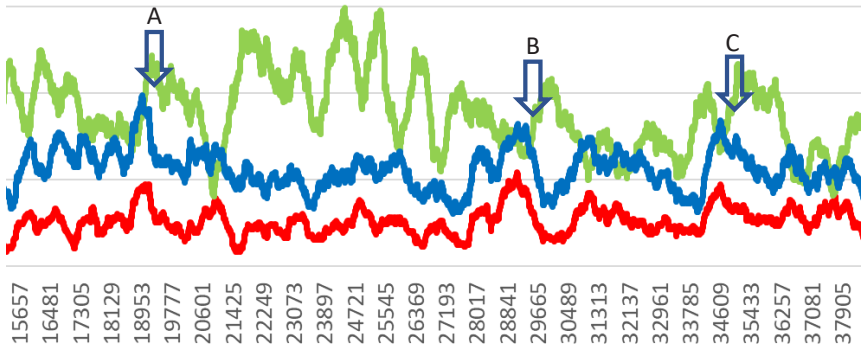
Figure 7: Peaks A, B, and C in BM (blue) new-word graph in comparison with hapax legomena (red) and repetition (green) graphs using scan window lengths of 500.

Figure 8: Peaks A, B, and C in BM (blue) and TEOM (red) new-word graphs using scan window lengths of 500 with values divided by scan window length (500)

Table 2: Peaks A, B, C and Text Alignment (words in bold are recurring new words; words in bold italics are hapax legomena)

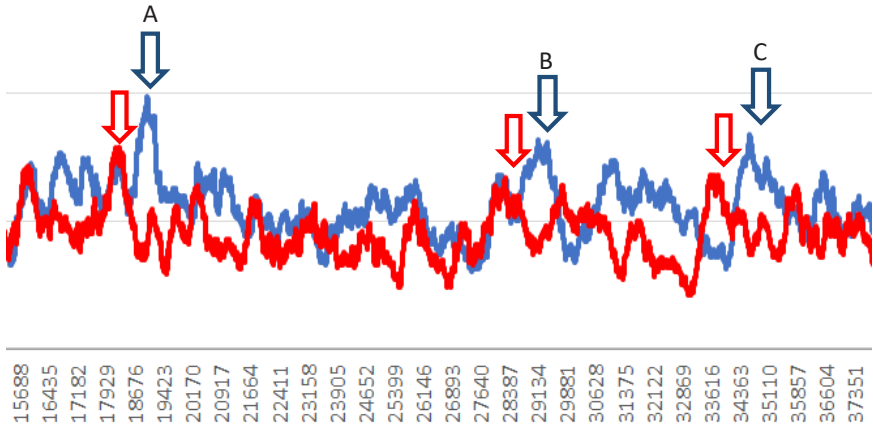| BM 18975-19474 (99)	TEOM 23177-23676 (66)
punya **seorang abang paiman dia lahir pada hari pasaran *paing*** maka dinamai dia dengan *suku* depan *pai* aku tiga tahun lebih muda dinamai **sanikem ayahku** bernama **sastrotomo** setelah kawin kata para tetangga nama itu **berarti jurutulis** yang **utama** kata orang **ayahku** seorang yang **rajin** ia **dihormati** karena satusatunya yang dapat **bacatulis** di desa baca tulis yang dipergunakan di kantor tapi ia tidak puas hanya jadi jurutulis ia impikan **jabatan** lebih tinggi sekali pun jabatannya sudah cukup tinggi dan terhormat ia tak perlu lagi *mencangkul* atau *meluku* atau **berkuli bertanam** atau **berpanen tebu** ayahku mempunyai banyak adik dan saudara *sepupu* sebagai jurutulis masih banyak kesulitan padanya untuk **memasukkan** mereka bekerja di pabrik jabatan lebih tinggi akan lebih **memudahkan** lagi pula akan semakin tinggi pada **pandangan** dunia apalagi ia ingin semua *kerabatnya* bisa bekerja di pabrik tidak sekedar jadi **kuli** dan **bawahan** paling rendah paling tidak *mandorlah* untuk membikin mereka jadi kuli tak perlu orang punya sanak jurutulis b semua orang bisa **diterima** jadi kuli kalau mandor setuju ia bekerja rajin dan semakin rajin lebih sepuluh tahun jabatan dan *pangkatnya* tak juga naik memang **gaji** dan **persen tahunan** selalu naik jadi *ditempuhnya* segala jalan dukun *jampi* **mantra bertirakat memutih berpuasa seninkamis** tak juga berhasil jabatan yang *diimpikannya* adalah **jurubayar kassier pemegang kas pabrikgula tulangan sidoarjo** dan siapa tidak *berurutan* dengan jurubayar pabrik paling sedikit *mandortebu* mereka datang untuk menerima uang dan *membubuhkan cap* **jempol** ia bisa menahan **upah** *mingguan* **kesatuan** si mandor kalau mereka menolak *cukaian* atas **penghasilan** para *kuhnya* sebagai jurubayar pabrik ia akan menjadi orang besar di tulangan pedagang akan membungkuk menghormati tuan tuan totok dan peranakan akan memberi tabik dalam melayu *guratan penanya* berarti uang ia akan termasuk golongan berkuasa dalam pabrik orang akan mendengarkan katanya tunggu di bangku situ untuk dapat menerima uang dan tangannya *mengibakan* bukan *kenaikan* jabatan kehormatan dan *ketakziman* yang ia dapatkan sebaliknya kebencian dan kejijikan orang dan jabatan **juru bayar** itu tetap tergantung *diawangawang tindakannya* yang *menjilat* dan **merugikan** orang *menjadikannya* **tersisih** dari pergaulan ia *terpencil* **ditengah** lingkungannya sendiri tapi ia tidak peduli ia memang keras hati kepercayaannya pada **kemurahan** dan **perlindungan** tuantuan kulit putih tak *terpatahkan* orang *muak* melihat usahanya menarik tuantuan belanda itu agar sudi datang ke rumah **seorangdua** memang datang juga dan *disugunya* dengan segala apa yang bisa menyenangkan mereka tapi jabatan itu tak juga tiba malah melalui dukun dan *tirakat* ia berusaha *menggendam* tuan administratur tuan besar kuasa agar sudi datang ke rumah juga tak berhasil sebaliknya ia sendiri sering berkunjung ke rumahnya bukan untuk **menemui pembesarnya** karena sesuatu urusan untuk membantu kerja di belakang tuan administratur tak pernah *mempedulikannya* aku sendiri merasa **risi** mendengar semua itu kadang dengan diamdiam kuperhatikan ayahku dan merasa **iba** betapa jiwa dan raganya *disesah* oleh impian itu betapa ia **hinakan** diri dan **martabat** sendiri tapi aku tak berani bicara apaapa memang kadang aku **berdoa** agar ia **menghentikan kelakuannya** yang **memalukan** itu para tetangga sering bilang lebih baik dan paling baik adalah **memohon** pada allah sampai *berapalah* kekuasaan manusia apalagi orang kulit putih pula **doaku**	**elder brother paiman he was born on the market day of *paing*** so he was **named** with the first *syllable pai* i was three years younger and named **sanikem** my father changed his name after he was married to **sastrotomo** the **neighbors** used to say the name meant the **foremost** *scribe* people said that my father was very **industrious** he was respected as the only person in village who could read and write the sort of reading and writing used in offices but he wasn't satisfied with just being a **clerk** in the factory he dreamed of a higher post even though the job he held was quite a respected one he no longer needed to *hoe* the ground or *plow* or *labor* or plant or harvest **sugar cane** my father had many younger brothers and sisters as well as *cousins* as a clerk he had great difficulty in getting them **jobs** at the factory a higher post would have made it **easier** and also it would have raised him up higher in the eyes of the world especially as he wanted his **relations** to be able to work in the factory as something more than just laborers and **coolies** at the very least they should be **foremen** you didn't need a blood **relative** as a clerk to get jobs as coolies **anybody** could get a job as a coolie as long as the **foreman agreed** he worked *diligently* and became even more *diligent* for more than ten years but still no **promotion** though his salary and **commission** rose every year so he tried every other way the traditional javanese magic men the *dukuns* magic *formulas* he even went on rice **fasts** *monday* and thursday fasts still no result he dreamed of becoming **paymaster cashier holder** of the *cash* of the **tulangan** sugar factory in **sidoardjo** and who did not have business with the factory paymaster there were the cane foremen they came to receive their money and leave their **thumbprints** if the foreman refused to accept a toll on the *coolies'* **wages** he could *withhold* the *foreman's gang's weekly* wages as paymaster he would be a big man in tulangan **merchants** would bow down in respect the pure and mixed-blood **tuans** would greet him in malay the **stroke** of his pen meant money he would be *counted* among the powerful in the factory people would listen to his words sit down on the bench there in order to receive their money from his hands **pathetic** these dreams did not bring him a rise in position respect or **esteem** on the contrary they brought hatred and disgust and the position of paymaster remained hanging in *limbo* far away his **crawling** behavior which often **harmed** his friends caused him to be cut off from society he was *isolated* in the midst of his own world but he didn't care he was indeed *hard-hearted* his trust in the *generosity* and **protection** of the white-skinned tuans could not be broken people were *sickened*

Table 2 (cont.): Peaks A, B, C and Text Alignment (words in bold are recurring new words; words in bold italics are hapax legomena)

BM 29333-29832 (81)	TEOM 36438-36939 (65)
Pada negeri dan bangsa ini sekali pun barang dua kali pernah disinggung dalam **diskusisekolah** tementeman menganggap bangsa ini **masih terlalu rendah untuk dibicarakan** secara **selintas** mereka *menyamaratakan* dengan pelacur **pelacurnya** yang memenuhi kembang **jepun warungwarung** kecil **restoran** dan *pangkas* rambut *verkoper* dan *kelontongnya* yang sama sekali tak dapat *mencerminkan* suatu pabrik yang menantang ilmu dan pengetahuan modern dalam suatu diskusisekolah waktu guruku tuan **lastendienst** mencoba menarik perhatian para siswa orang lebih banyak tinggal **mengobrol** pelan ia bilang di bidang ilmu jepang juga mengalami **kebangkitan kitasato** telah menemukan **kuman pes shiga** menemukan kuman *dysenteri* dan dengan demikian jepang telah juga berjasa pada ummat manusia ia *membandingkannya* dengan *sumbangan* bangsa belanda pada peradaban melihat aku mempunyai perhatian penuh dan membikin catatan lastendienst bertanya padaku dengan nada mendakwa eh minke **wakil** bangsa jawa dalam ruangan ini apa sudah di *sumbangkan* **bangsamu** pada ummat manusia bukan saja aku menggeragap mendapat pertanyaan dadakan itu **koleh** jadi seluruh dewa dalam **kotak** wayang ki *dalangjikan* hilang semangat hanya untuk menjawab maka jalan paling ampuh untuk tidak menjawab adalah *menyuarakan* kalimat ini ya meneer lastendienst sekarang ini *saya* belum bisa menjawab dan guruku itu menanggapi dengan senyum sangat manis itu sedikit *kutipan* dari catatanku tentang jepang dengan adanya tulisan dari majalah **pemberian** robert catatanku mendapatkan tambahan yang *lumayan* banyaknya tentang **kesibukan** di jepang untuk menentukan **strategi pertahanannya** aku tak banyak mengerti tentang hal demikian justru karena itu aku **catat** paling tidak akan menjadi bahan **bermegah** dalam diskusisekolah dikatakan adanya *persaingan* antara **angkatan** darat dengan angkatan laut jepang kemudian dipilih strategi *maritim* untuk pertahanannya dan angkatan darat dengan tradisi *samurainya* yang **berabad** merasa kurang senang *bagaimmana* tentang hindia belanda sendiri di dalamnya **dinyatakan** hindia belanda tidak mempunyai angkatan laut hanya angkatan darat yang **terdiri** dari *kepulauan* hindia belanda *setali* tiga uang mengapa kalau jepang **mengutamakan** laut hindia mengutamakan darat **bukankah** masalah pertahanan terhadap luar sama saja bukankah **jatuhnya** hindia belanda ke tangan inggris **nyaris seabad** yang lalu juga karena *lemahnya* angkatan laut di hindia mengapa itu tak **dijadikan** pelajaran dari majalah itu juga aku tahu hindia belanda tidak mempunyai angkatan laut kapal perang yang **mondarmandir** di hindia **bukanlah** milik hindia belanda tetapi milik kerajaan belanda *daendels* pernah membikin surabaya menjadi *pangkalan* angkatan laut pada masa hindia belanda tak punya *armada* satu pun nyaris seratus tahun setelah itu orang tak pernah memikirkan gunanya ada angkatan laut tersendiri untuk hindia tuantuan yang terhormat mempercayakan pertahanan laut inggris di **singapura** dan pertahanan laut amerika di *filipina* tulisan itu **membayangkan** sekiranya terjadi perang dengan jepang bagaimana akan halnya hindia belanda dengan **perairan** tak **terjaga** sedang angkatan laut kerajaan belanda hanya *kadangkadang* saja datang *meronda* **tidakkah** pengalaman tahun *1811* bisa berulang untuk **kerugian** belanda aku tak tahu apakah robert pernah membaca dan mempelajarinya sebagai pemuda yang ingin *berlanglang buana* sebagai pelaut boleh jadi ia telah mempelajarinya dan sebagai *pemuja* darah eropa kiranya dia *mengandalkan* **keunggulan ras** putih tulisan itu juga mengatakan jepang mencoba **meniru** inggris di perairan dan **pengarangnya** memperingatkan agar menghentikan ejekan terhadap bangsa itu sebagai monyet *peniru* pada setiap awal **pertumbuhan**	**as worthy of discussion** they *offhandedly equated* japan with the prostitutes who filled up the **kembang jepun** and with the little cafes *restaurants* and **barber shops** with the *hawker* and his goods none of these **reflected** the japan that was challenging modern science and learning in one discussion when my teacher mr **lastendienst** tried to get the students interested most just **chatted lazily** to each other he said that japan was also *experiencing* a *flowering* in the field of science *kitazato* had **discovered** the **plague** bacteria *shiga* had discovered *dysentery* bacteria and in that manner japan too had been of service to humanity he compared it with the dutch *nation's* **contribution** to civilization seeing that i was fully *engaged* in the subject and was taking notes mr lastendienst asked me in an accusing tone of voice eh minke the javanese *delegate* in this room what has your **nation contributed** to humanity i would not have been alone in being so startled to hear that sudden question in all *likelihood* all the gods in the chest of the shadow play *puppet-master* would have exhausted their **energy** just to answer so the best way of getting out of my difficulty was to *utter* the following sentence yes mr lastendienst i can't answer at this time and my teacher *reacted* to this with a sweet smile very sweet that's just a little from my notes about japan now with the articles in the magazine robert gave me my notes had been *supplemented* by quite a bit of **extra** information about the **current developments** in japan and the **struggle** over its defense **strategy** i didn't understand much about those things precisely because of that i noted it all down at the very least it would be excellent material for use in a school discussion it said there had been *competition* between the **japanese** army and navy a *maritime* defense strategy was then **chosen** and the army with its *centuries-old samurai* **tradition** was *dissatisfied* and the indies itself in the article it said the netherlands indies has no navy only an army japan is made up of *islands* the netherlands indies is just a great **string** of them why does japan **emphasize naval** defense while the indies emphasize the land isn't the problem of defense against the outside the same didn't the indies fall into the hands of the english a hundred years ago precisely because of the weakness of the indies navy why hasn't that **lesson** been learned the *warships* that *sailed* back and forth in indies waters did not *belong* to the netherlands indies but to the kingdom of the netherlands **governor-general** *daendels* had made surabaya a **naval** base in a period when he had not a single ship almost a hundred years later still no one gave any thought to the indies having its own navy the **honorable gentlemen** in **charge** put their trust in the **british** naval defenses of **singapore** and the american naval defenses of the *philippines* the article **speculated** about

Table 2 (cont.): Peaks A, B, C and Text Alignment (words in bold are recurring new words; words in bold italics are hapax legomena)

BM 34602-35101 (84)	TEOM 43657-44156 (72)
gamelan **nenenda** yang selalu ***terbungkus*** beledu merah bila tak ***ditabuh*** setiap tahun bukan hanya ***dilaras*** kembali juga ***dimandikan*** dengan air bunga **bersamaan** dengan gamelan datang juga ***juruluras*** ayahanda menghendaki bukan saja ***gamelannya*** juga ***larasnya*** harus **murni jawatimur** maka sejak pagi pendopo telah bising dengan bunyi orang **mengikir** dalam **melaras** pekerjaan ***administrasi*** kantor kebupatian b berhenti seluruhnya semua membantu tuan **niccolo moreno** seorang ***dekorator kenamaan*** yang didatangkan dari surabaya ia membawa serta kotak besar **alatalat** *hias* yang selama itu tak pernah kukenal dan pada waktu itu juga baru aku tahu **memajang** adalah satu **keahlian** tuan niccolo moreno datang atas **saran** tuan assisten residen b dibenarkan dan **ditanggung** oleh tuan residen surabaya pagi itu juga aku harus menemuinya dengan tangannya sendiri ia **ukur** tubuhku seperti hendak ***membikinkan*** pakaian untukku setelah itu ***dibiarkannya*** aku pergi pendopo itu telah ***dirubahnya*** menjadi ***arena*** dengan titik berat pada potret besar sri ratu wilhelmina dara cantik yang **pernah aku impikan dibawa dari surabaya** dilukis oleh seorang dengan nama jerman ***hüssenfeld*** aku masih tetap mengagumi kecantikannya **bendera** triwarna ***dipasang*** di manamana tunggal atau dua ***bersilang*** juga triwarna **pita** panjang ***berjuluran*** dari potret sri ratu ke seluruh pendopo dan bakalnya ***meraih*** para **hadirin** dengan **kewibawaannya** **tiangtiang** pendopo **dicat** dengan ***cat tepung*** yang baru kuperhatikan waktu itu pula dan dapat kering dalam hanya dua jam daun ***beringin*** dan **janur** kuning dalam **keserasian** warna tradisi mengubah dinding dan tiangtiang yang ***keringkerontang*** menjadi sejuk dan memaksa orang untuk menikmati dengan ***pengelihatannya*** maka mata pun ***diayunkan*** oleh permainan warna bungabungaan kuning biru merah putih dan **ungu** indah **meresap** bungabungaan yang dalam kehidupan seharihari **berpisahan** dan dengan diamdiam ***berjengukan*** pada pagar malam kebesaran dalam hidup ayahanda tiba juga gamelan sudah lama ***mendayu dayu*** pelahan tuan niccolo moreno sibuk dalam kamarku **merias** aku siapa pernah sangka aku yang sudah dewasa ini pula seakan aku dara akan naik ke **puadai pengantin** selama merias tak ***hentinya*** ia bicara dalam belanda yang kedengaran aneh ***datar*** seperti keluar dari rongga mulut pribumi jelas ia bukan belanda menurut ceritanya ia sering merias para bupati termasuk ayahku sekarang ini para raja di jawa dan **sultan** di **sumatra** dan ***borneo*** ia telah banyak membikin rencana pakaian mereka dan masih tetap dipergunakan sampai sekarang katanya pula pakaian pasukan pengawal para raja di jawa ia juga yang ***merencanakan*** diamdiam aku mendengarkan tidak mengiakan juga tidak membantah sekali pun tak percaya sepenuhnya ia telah kenakan padaku **kemejadada** berenda kaku seperti terbuat dari selembar kulit penyu tak mungkin rasanya ***membongkok*** dengan kemejadada ini ***gombaknya*** yang kaku seperti kulit sapi juga membikin leher segan untuk maksudnya supaya badan tetap tegap tidak sering menoleh pandang lurus seperti gentlemen sejati kemudian ia kenakan padaku kain batik dengan **ikat** pinggang perak gaya ***pengenaan*** kain itu diatur **sedemikian rupa** sehingga muncul **watak** ke ***jawatimurannya*** yang gagah itu yang kiranya dikehendaki ayah aku tetap **manda** seperti anak dara sebuah **blangkon** dengan gaya ***perpaduan*** antara jawatimur dan madura sama sekali baru ***kreasi*** niccolo moreno sendiri terpasang pada kepalaku menyusul sebilah **keris** ***bertatahkan*** **permata** kemudian baju **lasting** hitam berbentuk jas pendek dengan ***cowak*** pada	**had once dreamed after brought from surabaya** the work of a german artist named ***hüssenfeld*** i still admired her beauty the dutch **tricolors** were hung everywhere *singly* or in *twos* tricolor **ribbon** also streamed out from the portrait to all parts of the pavilion and would later ***captivate*** the audience with its authority the ***pavilion's*** columns were painted with some new kind of paint made from ***flour*** that ***dried*** within two hours **banyan-tree** leaves and ***greenish-yellow*** coconut ***fronds*** in traditional **color** *harmonies* **transformed** the dry barren walls into something **refreshing** and ***impelled*** people to enjoy their beauty eyes were **drawn** by the play of ***flowers'*** colors yellow blue red white and **purple** a *saturating* beauty flowers that in day-to-day life stuck **separately** and silently out along ***fences*** the big night in my father's life arrived the gamelan had already been ***rumbling*** softly and slowly for some time mr niccolo moreno was busy in my room dressing me up and ***adorning*** me who would have ever guessed that i already an adult would be dressed up by somebody else a white person too as if i were a maiden about to **ascend** the wedding throne all the time he was dressing me he spoke in a ***strangesounding monotone*** dutch as if it came out of the chest of a native he obviously wasn't dutch according to his story he often dressed and adorned the bupatis including my father tonight and the **sultans** of **sumatra** and **borneo** he'd **designed** many of their clothes and even now was often summoned by them he said also that the ***costumes*** of the ***guards*** of the kings of java were designed by him silently i listened to his stories neither **affirming** nor refuting them although i didn't believe them fully either he had dressed me in an embroidered ***vest* stiff** as if made from ***tortoiseshell*** i could never have bent over in it the stiff leather **collar** ***dissuaded*** my neck from turning around indeed the **intent** was that my body should be straight and stiff not turning around ***frequently*** eyes straight ahead like a true ***gentleman*** then a batik sarong with a silver **belt** the **style** in which the batik was **worn** truly brought out that dashing east javanese character that's what father no doubt wanted i **suffered** all this like a young maiden a batik **blangkon** ***headdress*** a mixture of east javanese and madurese styles something entirely new niccolo ***moreno's*** own **creation** was placed upon my head then came a **ceremonial** ***sheathed*** short sword a **keris** ***inlaid*** with jewels then a black **outer** **upper** ***garment*** like a coat with a cut at the back so the people could admire the beauty of my keris a bow **tie** made my neck usually ***active*** guiding my eyes to their targets feel as if it were being ***snared*** hot ***perspiration*** began to **soak** my back and chest in the mirror i found myself looking like a **victorious** *knight* out of those stories of the ***legendary eleventh-century*** prince panji

Table 3: BM Peak A Seed-words (numbers in brackets refer to number of appearances in the scan window; numbers in parentheses refer to number of subsequent appearances)

BM 18975-19474 (99)

	1	2-5	6-20	Top 5 (seed-words)
dan [21]; ia [17]; orang [9]; tak [9]; yang [9]; akan [7]; di [7]; itu [7]; lebih [6]; untuk [6]; jadi [5]; juga [5]; mereka [5]; pabrik [5]; tidak [5]; Tuan [5]; aku [4]; bisa [4]; datang [4]; dengan [4]; memang [4]; pada [4]; paling [4]; sendiri [4]; tapi [4]; tinggi [4]; agar [3]; atau [3]; bekerja [3]; ke [3]; para [3]; semua [3]; uang [3]; adalah [2]; administratur [2]; apalagi [2]; baik [2]; banyak [2]; berhasil [2]; besar [2]; betapa [2]; bukan [2]; dalam [2]; dapat [2]; dia [2]; dinamai [2]; dukun [2]; kadang [2]; kalau [2]; karena [2]; kata [2]; kulit [2]; lagi [2]; mandor [2]; menerima [2]; merasa [2]; naik [2]; perlu [2]; pula [2]; punya [2]; putih [2]; rumah [2]; sebagai [2]; sebaliknya [2]; segala [2]; semakin [2]; seorang [2]; sering [2]; sudi [2]; tahun [2]; tetangga [2]; tuantuan [2]; abang [1]; adik [1]; Allah [1]; apa [1]; apaapa [1]; atas [1]; baca [1]; bangku [1]; belakang [1]; Belanda [1]; berani [1]; berkuasa [1]; berkunjung [1]; bernama [1]; bertirakat [1]; berusaha [1]; bicara [1]; bilang [1]; cukup [1]; dapatkan [1]; dari [1]; depan [1]; desa [1]; diamdiam [1]; dipergunakan [1]; diri [1]; dunia [1]; golongan [1]; hanya [1]; hari [1]; hati [1]; impian [1]; impikan [1]; ingin [1]; jabatannya [1]; jalan [1]; jiwa [1]; kantor [1]; katanya [1]; kawin [1]; kebencian [1]; kehormatan [1]; kejijikan [1]; kekuasaan [1]; kepercayaannya [1]; keras [1]; kerja [1]; kesulitan [1]; kuasa [1]; kuperhatikan [1]; lahir [1]; lingkungannya [1]; maka [1]; malah [1]; manusia [1]; masih [1]; melalui [1]; melayu [1]; melihat [1]; membantu [1]; memberi [1]; membikin [1]; membungkuk [1]; mempunyai [1]; memutih [1]; menahan [1]; menarik [1]; mendengar [1]; mendengarkan [1]; menghormati [1]; menjadi [1]; menolak [1]; menyenangkan [1]; muda [1]; nama [1]; oleh [1]; padanya [1]; pedagang [1]; peduli [1]; peranakan [1]; pergaulan [1]; pernah [1]; puas [1]; pun [1]; raganya [1]; rendah [1]; rumahnya [1]; sampai [1]; sanak [1]; satusatunya [1]; saudara [1]; sedikit [1]; sekali [1]; sekedar [1]; selalu [1]; sepuluh [1]; sesuatu [1]; setelah [1]; setuju [1]; si [1]; siapa [1]; situ [1]; sudah [1]; tabik [1]; tangannya [1]; tergantung [1]; terhormat [1]; termasuk [1]; tetap [1]; tiba [1]; tiga [1]; totok [1]; tulis [1]; tunggu [1]; urusan [1]; usahanya [1]	berapalah (1) [1]; berkuli (1) [1]; berpanen (1) [1]; berpuasa (1) [1]; bertanam (1) [1]; berurutan (1) [1]; cap (1) [1]; cukaian (1) [1]; diawangawang (1) [1]; diimpikannya (1) [1]; disesah (1) [1]; disugunya (1) [1]; ditempuhnya (1) [1]; guratan (1) [1]; jampi (1) [1]; kenaikan (1) [1]; kerabatnya (1) [1]; ketakziman (1) [1]; kuhnya (1) [1]; mandorlah (1) [1]; mandortebu (1) [1]; Meluku (1) [1]; membubuhkan (1) [1]; mempedulikannya (1) [1]; mencangkul (1) [1]; menggendam (1) [1]; menjilat (1) [1]; mingguan (1) [1]; muak (1) [1]; pabrikgula (1) [1]; Pai (1) [1]; Paing (1) [1]; pangkatnya (1) [1]; penanya (1) [1]; penghasilan (1) [1]; seninkamis (1) [1]; sepupu (1) [1]; suku (1) [1]; tebu (1) [1]; terpatahkan (1) [1]; terpencil (1) [1]; tersisih (1) [1]; tindakannya (1) [1]; tirakat (1) [1]	diterima (2) [1]; gaji (2) [1]; tahunan (2) [1]; mantra (2) [1]; kassier (2) [1]; kas (2) [1]; jempol (2) [1]; menjadikannya (2) [1]; ditengah (2) [1]; kemurahan (2) [1]; seorangdua (2) [1]; pembesarnya (2) [1]; risi (2) [1]; hinakan (2) [1]; martabat (2) [1]; menghentikan (2) [1]; doaku (2) [1]; Paiman (3) [1]; pasaran (3) [1]; pemegang (3) [1]; kesatuan (3) [1]; bayar (3) [1]; perlindungan (3) [1]; memudahkan (4) [1]; bawahan (4) [1]; persen (4) [1]; Sidoarjo (4) [1]; upah (4) [1]; juru (4) [1]; merugikan (4) [1]; iba (4) [1]; berdoa (4) [1]; kelakuannya (4) [1]; utama (5) [1]; kuli (5) [3]; jurubayar (5) [3]	rajin (6) [3]; bacatulis (6) [1]; memohon (6) [1]; dihormati (7) [1]; memasukkan (7) [1]; pandangan (7) [1]; menemui (8) [1]; mengibakan (9) [1]; Tulangan (10) [2]; memalukan (12) [1]; Sastrotomo (14) [1]; jabatan (14) [7]; jurutulis (15) [4]; Sanikem (16) [1]; berarti (16) [2]; ayahku (19) [4]	Sastrotomo (14) [1]; jurutulis (15) [4]; Sanikem (16) [1]; berarti (16) [2]; ayahku (19) [4]

Table 3 (cont.): TOEM Peak A Seed-words (numbers in brackets refer to number of appearances in the scan window; numbers in parentheses refer to number of subsequent appearances)

TOEM 23177-23676 (66)

	1	2-5	6-20	Top 5 (seed-words)
the [35]; he [20]; in [16]; and [15]; a [12]; as [12]; of [11]; to [11]; his [10]; was [9]; would [7]; be [6]; factory [6]; on [5]; or [5]; could [4]; him [4]; but [3]; even [3]; father [3]; have [3]; higher [3]; money [3]; more [3]; my [3]; no [3]; not [3]; people [3]; their [3]; they [3]; with [3]; at [2]; blood [2]; did [2]; didn't [2]; down [2]; dreamed [2]; every [2]; from [2]; get [2]; had [2]; it [2]; job [2]; just [2]; magic [2]; meant [2]; name [2]; position [2]; post [2]; receive [2]; respect [2]; respected [2]; s [2] ; so [2]; still [2]; than [2]; there [2]; though [2]; used [2]; very [2]; were [2]; who [2]; world [2]; years [2]; younger [2]; able [1]; accept [1]; after [1]; also [1]; among [1]; away [1]; became [1]; becoming [1]; behavior [1]; being [1]; bench [1]; big [1]; born [1]; bow [1]; bring [1]; broken [1]; brother [1]; brothers [1]; brought [1]; business [1]; came [1]; care [1]; caused [1]; changed [1]; contrary [1]; coolie [1]; cut [1]; day [1]; difficulty [1]; disgust [1]; dreams [1]; elder [1]; especially [1]; eyes [1]; far [1]; first [1]; for [1]; friends [1]; getting [1]; great [1]; greet [1]; ground [1]; hands [1]; hanging [1]; harvest [1]; hatred [1]; held [1]; I [1]; if [1]; indeed [1]; Javanese [1]; laborers [1]; least [1]; leave [1]; listen [1]; long [1]; longer [1]; made [1]; Malay [1]; man [1]; many [1]; married [1]; men [1]; midst [1]; mixed [1]; need [1]; needed [1]; off [1]; offices [1]; often [1]; one [1]; only [1]; order [1]; other [1]; own [1]; pen [1]; person [1]; plant [1]; powerful [1]; pure [1]; quite [1]; raised [1]; read [1]; reading [1]; refused [1]; remained [1]; result [1]; rice [1]; rise [1]; said [1]; salary [1]; satisfied [1]; say [1]; should [1]; sisters [1]; sit [1]; skinned [1]; society [1]; something [1]; sort [1]; ten [1]; that [1]; them [1]; these [1]; three [1]; thursday [1]; traditional [1]; tried [1]; trust [1]; up [1]; village [1]; wanted [1]; wasn't [1]; way [1]; well [1]; went [1]; which [1]; white [1]; words [1]; work [1]; worked [1]; write [1]; writing [1]; year [1]; you [1]	cash (1) [1] commission (1) [1] coolies' (1) [1] counted (1) [1] cousins (1) [1] diligent (1) [1] diligently (1) [1] dukuns (1) [1] foreman's (1)[1] formulas (1) [1] gang's (1) [1] generosity (1) [1] hard-hearted (1) [1] hoe (1) [1] holder (1) [1] isolated (1) [1] limbo (1) [1] merchants (1) [1] Monday (1) [1] Pai (1) [1] Paing (1) [1] plow (1) [1] promotion (1) [1] scribe (1) [1] sickened (1) [1] syllable (1) [1] toll (1) [1] weekly (1) [1] withhold (1) [1]	market (2) [1] foremost (2) [1] industrious (2) [1] coolies (2) [2] relative (2) [1] fasts (2) [2] thumbprints (2) [1] stroke (2) [1] esteem (2) [1] harmed (2) [1] foreman (2) [2] Paiman (3) [1] labor (3) [1] sugar (3) [2] cane (3) [2] jobs (3) [2] foremen (3) [2] agreed (3) [1] tuans (3) [2] pathetic (3) [1] crawling (3) [1] protection (3) [1] cashier (4) [1] Sidoardjo (4) [1] easier (5) [1] wages (5) [2]	anybody (6) [1] paymaster (6) [4] named (7) [2] neighbors (8) [1] Tulangan (10) [2] clerk (11) [3] rose (11) [1] relations (14) [1] Sastrotomo (14) [1] Sanikem (18) [1]	Tulangan (10) [2] clerk (11) [3] Sastrotomo (14) [1] relations (14) [1] Sanikem (18) [1]

"**Ayahku** bernama **Sastrotomo** setelah kawin. Kata para tetangga, nama itu berarti: **Jurutulis** yang utama."

(My father took on the name Sastrotomo after marriage. The neighbors said this name meant: the first among scribes.)

Family & Community

ayahku (19) [4] {my father}
Sanikem (16) [1] (proper name)
Sastrotomo (14) [1] (proper name)
Paiman (3) [1] (proper name)
kerabatnya (1) [1] (relative, family)
Pai (1) [1] (proper name)
Paing (1) [1] (proper name)
sepupu (1) [1] {cousin}
suku (1) [1] {clan, ethnicity}
dinamai (0) [2] {named}
tetangga (0) [2] {neighbour}
abang (0) [1] {elder brother}
adik (0) [1] {younger brother or sister}
nama (0) [1] {name}
sanak (0) [1] {relative}
saudara (0) [1] {brother}

Honor

memalukan (12) [1] {shameful, disgraceful}
dihormati (7) [1] {honored}
pandangan (dunia) (7) [1] {public opinion}
hinakan [diri] (2) [1] {to lower oneself}
martabat {sendiri} (2) [1] {one's own honor}
ketakziman (1) [1] {honor and respect}
kehormatan (0) [1] {honor}
menghormati (0) [1] {to honor}
tabik (0) [1] {greeting}
terhormat (0) [1] {honored}
golongan (0) [1] {group, class}

Employment and Status

jabatan {lebih tinggi} (14) [7] {higher post or position}
pangkatnya {tak naik} (1) [1] {he did not rise in rank}
kenaikan (1) [1] {promotion}

Plantation and Factory Labor

jurutulis (15) [4] {scribe}; Tulangan (10) [2] {place name}; memasukkan (7) [1] {enter employment}; rajin (6) [3] {diligent}; bacatulis (6) [1] {read-write}; kuli (5) [3] (laborer); jurubayar (5) [3] {paymaster}; mengikan (4) [1] {financial loss}; bawahan (4) [1] {subordinate}; persen (4) [1] {gratuity}; Sidoarjo (4) [1] {place name}; upah (4) [1] {pay}; juru (4) [1] {skilled worker}; pasaran (3) [1] {market}; bayar (3) [1] {to pay}; gaji (2) [1] {wage}; kassier (2) [1] {cashier}; kas (2) [1] {cash}; jempol (2) [1] {thumbprint}; berkuli (1) [1] {to work as a laborer}; berpanen (1) [1] {to plant}; cukaian (1) [1] {toll, tariff}; guratan {pena} (1) [1] {stroke of a pen}; kulinya (1) [1] {his laborer}; mandoriah (1) [1] {oversee}; mandortebu (1) [1] {foreman in a sugarcane plantation}; mencangkul (1) [1] {to hoe}; pabrikgula (1) [1] {sugar factory}; penanya (1) [1] {his pen}; tebu (1) [1] {sugarcane}; pabrik (0) [5] {factory}; Tuan (0) [5] {Mr. for Westerners}; uang (0) [3] {money}; administratur (0) [2] {administrator}; mandor (0) [2] {to oversee}; tuantuan (0) [2] {plural of "Tuan"}; baca (0) [1] {read}; Belanda (0) [1] {Netherlands}; jabatannya (0) [1]; kantor (0) [1] {office}; kerja (0) [1] {work}; pedagang (0) [1] {trader, merchant}; tulis (0) [1] {write}

"Ia impikan **jabatan** lebih tinggi, sekali pun jabatannya sudah cukup tinggi dan **terhormat.**"

(He dreamed of a higher post even though his position was already high enough and honored.)

Magic and Religion

mantra (2) [1] {mantra}
doaku (2) [1] {my prayer}
berdoa (4) [1] {to pray}
berpuasa (1) [1] {to fast}
jampi (1) [1] {spell}
dukun [2] {traditional holy man}
Allah [1]

Figure 9: BM Major Themes and Seed-Words

"My **father** changed his name after he was married to **Sastrotomo**. The neighbors used to say the name meant the foremost **scribe**."

Plantation and Factory Labor

clerk (11) [3]; paymaster (6) [4]; Tulangan (10) [2]; market (2) [1]; industrious (2) [1]; coolies (2) [2]; thumbprints (2) [1]; foreman (2) [2]; labor (3) [1]; sugar (3) [2]; cane (3) [2]; jobs (3) [2]; foremen (3) [2]; tuans (3) [2]; cashier (4) [1]; Sidoardjo (4) [1]; wages (5) [2]; cash (5) [1]; coolies' (1) [1]; diligent (1) [1]; diligently (1) [1]; foreman's (1) [1]; gang's (1) [1]; hoe (1) [1]; merchants (1) [1]; plow (1) [1]; scribe (1) [1]; toll (1) [1]; factory (0) [6]; money (0) [3]; job (0) [0]; [2]; coolie (0) [1]; harvest (0) [1]; laborers (0) [1]; offices (0) [1]; pen (0) [1]; read (0) [1]; reading (0) [1]; salary (0) [1]; society (0) [1]; words (0) [1]; work (0) [1]; worked (0) [1]; write (0) [1]; writing (0) [1]

Employment and Status

commission (1) [1]
promotion (1) [1]
position (0) [2]
post (0) [2]

"He dreamed of a higher **post** even though the job he held was quite a **respected** one."

Honor

esteem (2) [1]
pathetic (3) [1]
crawling (3) [1]
respect (0) [2]
respected (0) [2]

(outside the scan window)
humiliated (x) [10]
dignity (x) [4]

Family & Community

neighbors (8) [1]
relations (14) [1]
Sastrotomo (14) [1]
Sanikem (18) [1]
named (7) [2]
relative (2) [1]
Paiman (3) [1]
cousins (1) [1]
Pai (1) [1]
Paing (1) [1]
father (0) [3]
blood (0) [2]
brother (0) [1]
brothers (0) [1]

Magic and Religion

fasts (2) [2]
dukuns (1) [1]
formulas (1) [1]
magic (0) [2]

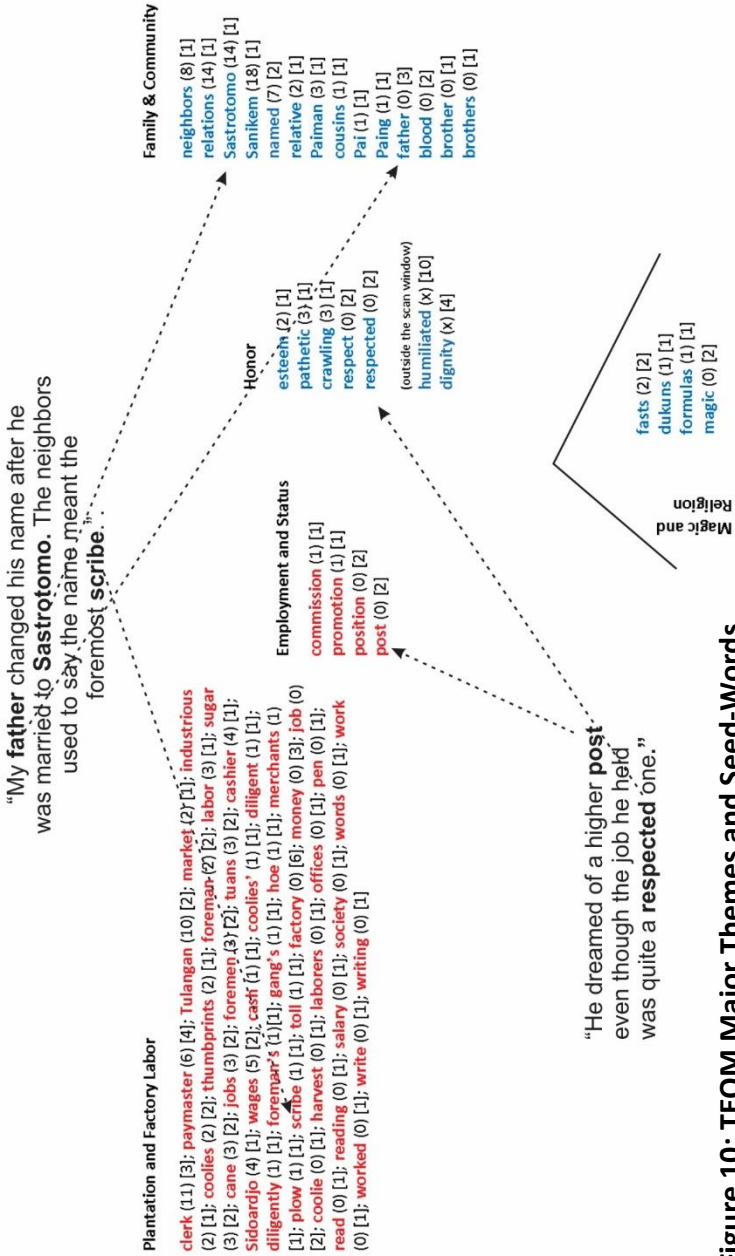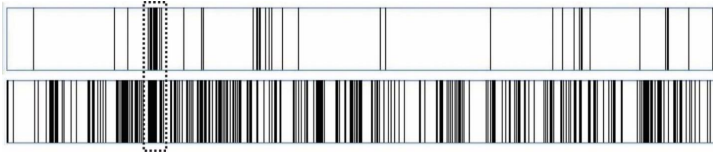**Figure 10: TEOM Major Themes and Seed-Words**

Figure 11: Concordance plots of textual occurrences of "family & kinship" (top) and "economic" (bottom) terms in BM. Points 18975-19474 indicated in box with dotted lines.

Table 4: Reduction in TEOM of "Honor" to "Respect"

BM		TEOM
Ke(hormat)an	Respect; Esteem; Honor	Respect
Ke(takzim)an		Esteem
Martabat	Rank; Status; Prestige; Dignity	Dignity

Appendix A: textanalysis-1 and textanalysis-2 Programs (the symbol ">" at line beginnings should be replaced by a space when encoding)

```
##########################################################
# PROGRAM NAME: textanalysis-1.py
# AUTHOR: R.G.
# INSTITUTION: UNIVERSITY OF THE PHILIPPINES DILIMAN
# DATE: June 2017
# WRITTEN IN PYTHON 3.6
##########################################################
# SCAN LENGTH
# ASSIGN A VALUE FOR THE LENGTH OF THE SCAN WINDOW
# RECOMMENDED VALUES DEPENDING ON THE LENGTH OF THE TEXT:
# 20, 50, 100, 300, 500, 1000
SCN=[INPUT SCAN LENGTH HERE]
filename="[INPUT FILENAME HERE]"
##########################################################

print ('RUNNING TEXT ANALYSIS ')
text=open(filename + '.txt', 'r')
S=text.read()
L0 = S.split( ); TITLE=L0[0]+'-'+L0[1]+'-'+L0[2]
L1 = L0
L2 = [0]*len(L0)
L3 = [0]*(len(L0)-SCN)
L4 = [0]*(len(L0)-SCN)
```

```
L5 = S.split( )
typs = 0
for i in range (len(L1)):
>>>>for j in range (len(L1[i:])):
>>>>>>>>if L1[i]=='-':
>>>>>>>>>>>>continue
>>>>>>>>if i==j+i and L1[i] == L1[j+i]:L2[i]=1; continue
>>>>>>>>if L1[i] == L1[j+i]:L2[i]+=1; L1[j+i]='-';typs+=1
for i in range (len(L1)-SCN):
>>>>for j in range (SCN):
>>>>>>>>>>if L2[j+i]==1:L3[i]+=1
>>>>>>>>>>if L2[j+i]>=1:L4[i]+=1
>>>>>>>>>>if j+i==len(L2)-1: break
count=1
textdata1=open('1-'+'TEXT-'+TITLE +'.txt', 'w')
textdata1.write(TITLE+'\n'+'TOKENS: '+str(len(L0))+'   TYPES: '+ str(len(L0) - typs))
textdata1.write('   LEXICAL COMPLEXITY (TYPES/TOKENS): '+ str((len(L0)-typs)*1.0/
len(L0))+'\n\n')
textdata1.write('WORDLIST IN SEQUENCE' + '\n\n')
for b in range (len(L5)):
>textdata1.write(str(count)+'\t'+ L5[b] + '\n')
>count+=1
textdata1.close( )
count=1
zip (L0,L1)
textdata2=open('2-'+'WORDFREQS-'+str(SCN)+'-'+TITLE+'.txt', 'w')
textdata2.write(TITLE+'\n'+'NEW WORDS IN SEQUENCE AND FREQUENCIES' + '\n\n')
for a,b in zip(L1,L2):
>textdata2.write(str(count)+'\t'+ a + '\t\t' + str(b)+'\n')
>count+=1
textdata2.close( )
textdata3=open('3-'+'NEWWORDS-'+str(SCN)+'-'+TITLE+'.csv', 'w')
ctr=1
for a in L4[:]:
>>>>textdata3.write(str(ctr)+', '+str(a) + ', '+ str(a/SCN) +'\n')
>>>>ctr=ctr+1
textdata3.close( )
textdata4=open('4-'+'HAPAXA-'+str(SCN)+'-'+TITLE+'.csv', 'w')
ctr=1
for a in L3[:]:
>>>>textdata4.write(str(ctr)+', '+str(a) + ', '+ str(a/SCN) +'\n')
>>>>ctr=ctr+1
textdata4.close( )
print ('FINISHED!')

#########################################################
# PROGRAM NAME: textanalysis-2.py
# AUTHOR: R.G.
# INSTITUTION: UNIVERSITY OF THE PHILIPPINES DILIMAN
# DATE: June 2017
# WRITTEN IN PYTHON 3.6 (ANACONDA)
# REQUIRES INSTALLATION OF NETWORKX
#########################################################
```

```
# SCAN LENGTH
# ASSIGN A VALUE FOR THE LENGTH OF THE SCAN WINDOW
# RECOMMENDED VALUES DEPENDING ON THE LENGTH OF THE TEXT:
# 20, 50, 100, 300, 500, 1000
SCN=[INPUT SCAN LENGTH HERE]
filename="[INPUT FILENAME HERE]"
##########################################################

###########
#REPETITION
###########

print ('RUNNING TEXT ANALYSIS ')
text=open(filename + '.txt', 'r')
SS=text.read();
S=' '+SS+' '
L0=S.split( )
TITLE=L0[0]+'-'+L0[1]+'-'+L0[2]
text1=open('stoplist.txt', 'r')
STP=text1.read();
STP=STP.split( )
for a in range(len(STP)):
>>>>stpwrd=' '+STP[a]+' '
>>>>S=S.replace(stpwrd,' @@@ ')
L1 = L0
L2 = [0]*len(L0)
L3 = [0]*(len(L0)-SCN)
L4 = [0]*(len(L0)-SCN)
for i in range (len(L1)):
>>>>for j in range (len(L1[i:])):
>>>>>>>>if L1[i]=='@@@':
>>>>>>>>>>>>continue
>>>>>>>>if i==j+i and L1[i] == L1[j+i]:L2[i]=1; continue
>>>>>>>>if L1[i] == L1[j+i]:L2[i]+=1; L1[j+i]='@@@'
FRQ={}
for a,b in zip(L1,L2):
>>>>FRQ[a]=b
textdata5=open('5-'+'REPETITION-'+str(SCN)+'-'+ TITLE + '.csv', 'w')
L0=S.split( )
ctr=1
for i in range (len(L0)-SCN):
>>>>freq=0
>>>>for j in range (SCN):
>>>>>>>>if i+j >len(L0)-1: break
>>>>>>>>freq=freq+FRQ[L0[i+j]]
>>>>textdata5.write(str(ctr)+', '+str(freq) + ', '+ str(freq/(len(L0)-SCN)) + '\n')
>>>>ctr=ctr+1
text.close ()
text1.close ()
textdata5.close ()
print ('FINISHED REPETITION!')
```

```
####################
# LEXICAL COMPLEXITY
####################

S=' '+SS+' '
L0=S.split( )
L1 = [0]*len(L0)
ctr=1
textdata6=open('6-'+'LEXCOMP-'+str(SCN)+'-'+ TITLE + '.csv', 'w')
for i in range (len(L0)-SCN):
>>>>L2=[0]*SCN
>>>>for j in range (SCN):
>>>>>>>>L2[j]=L0[i+j]
>>>>L3=L2
>>>>L4=' '.join(L3)
>>>>for k in range (SCN):
>>>>>>>>if L3[k]=='@@@': continue
>>>>>>>>if L3[k]=='XXX': continue
>>>>>>>>wrd =' '+ L3[k]+' '
>>>>>>>>L4=L4.replace(wrd,' @@@ ')
>>>>>>>>L3=L4.split( )
>>>>>>>>L3[k]='XXX'
>>>>>>>>L4=' '.join(L3)
>>>>L3=L4.split( )
>>>>for l in range (SCN):
>>>>>>>>if L3[l]=='XXX':
>>>>>>>>>>>>>>>>L1[i]+=1
for l in range (len(L0)-SCN):
>>>>textdata6.write(str(ctr)+', '+str(L1[l]/SCN) + '\n')
>>>>ctr=ctr+1
textdata6.close ()
print ('FINISHED LEXICAL COMPLEXITY!')

########################
# BETWEENNESS CENTRALITY
########################

S=' '+ SS + ' '
L0=S.split( )
L1=['0']*len(L0)
L2=['0']*len(L0)
L3=['0']*len(L0)
L4=['0']*len(L0)
for i in range (len(L0)-1):
>>>>L1[i]=L0[i]
>>>>L2[i]=L0[i+1]
import networkx as nx
G=nx.Graph ()
for i in range (len(L1)):
>>>>G.add_edge (L1[i], L2[i])
S= nx.betweenness_centrality (G)
for j in range (len(L0)-1):
>>>>L4[j]=(S[L1[j]])
```

```
textdata7=open('7-'+'BETCEN-'+str(SCN)+'-'+TITLE +'.csv', 'w')
ctr=1
for i in range (len(L0)-SCN):
>>>>z=0.0
>>>>for j in range (SCN):
>>>>>>>>if i+j >len(L0)-1: break
>>>>>>>>z= float(L4[i+j])+z
>>>>textdata7.write(str(ctr)+', '+str(z/SCN)+'\n')
>>>>ctr=ctr+1
textdata8=open('8-'+'BETCENLIST-'+str(SCN)+'-'+TITLE +'.csv', 'w')
ctr=1
for i in range (len(L0)):
>>>>textdata8.write(str(ctr)+', '+L0[i]+', '+ str(S[L0[i]]) +'\n')
>>>>ctr=ctr+1
textdata7.close ()
textdata8.close ()

print ('FINISHED BETWEENNESS CENTRALITY!')
```