A Javanese Corpus: Its Use in Maintaining Javanese As A Vibrant, Relevant, Living Language

Allan F. Lauder
University of Indonesia - UNITED KINGDOM
allan.lauder@gmail.com

Abstract

The Javanese language is an important world language with a rich literary heritage and despite being an indigenous language in Indonesia, the term 'minority' hardly seems appropriate. However, despite the large number of speakers of Javanese, the language is not immune to the processes of language change. Javanese is a linguistic and cultural treasure. It is one of the world's major languages. Yet its status as a regional language puts and its relationship with Indonesian are not helping it continue as a vibrant, relevant living language. In order to continue as the language which Javanese remain proud of, it must adapt to social changes and find ways to counter potential or emerging threats that could push it from its premier position into decline. The development of corpora is not a small undertaking and considerable resources and time are required. However, a Javanese corpus would be a treasure that is worthy of the language and the culture it embodies, and it would be much appreciated by academics, policy makers and writers.

Keywords: Javanese corpus, maintaining, living language, and indigenous language

A. Introduction: Why Create a Corpus of Javanese?

The regional indigenous languages of Indonesia (also referred to variously as vernaculars or provincial languages) are part of a complex linguistic situation that is generally seen as comprised of three categories: Indonesian (Bahasa Indonesia), the regional indigenous languages, and foreign languages (Alwi and Sugono, 2000), (Renandya, 2000: 115). Thirteen of the regional languages have a million or more speakers, accounting for 69.91% of the total population (Lauder, 2004b: 3-4).

The Javanese language is the most influential indigenous regional language in Indonesia. With over75 million speakers, it is the most populous with more speakers than Italian. It is the first of the estimated 726 extant regional languages in the country (Grimes, 2000, Lewis, 2009). Javanese, like all human languages, is subject to language change. Contemporary Javanese is different from Javanese at different periods in a history that goes back to the 9th century.

There is enough anecdotal evidence to suspect that Javanese today is being used less, giving way in places to the use of Indonesian. A possible shift from Javanese to Indonesian can be inferred from the national census which shows that Indonesian is being understood and used by more and more people. It is, however, difficult to get hard, reliable data on the use of Javanese itself in order to know to what extent it is occurring (if it is) or what the details of this are. It is also possible to find writers to point to a reduction in the use of the more refined krama level of Javanese, and who see a trend whereby the complex system of distinct social levels of the language are moving towards a simplification (Wahab, 2006). The evidence while incidental would be in line with general global trends of language loss.

These processes can be seen from the perspective of language change. They can also be seen from the perspective of the status of Javanese as a regional indigenous language, alongside the national language, Indonesian, and foreign languages, in particular English.

As such, the Javanese language can be considered in the context of the rights of indigenous languages to preservation based on a philosophical and legal position of the rights of minorities to preservation and protection. See, for example, Lauder (2007) on the Indonesian regional languages and Burch (2009) for the picture in Europe. The idea that indigenous languages have rights and should be preserved is rooted in the idea of the value of diversity (Skutnabb-Kangas, 2000).

Because Javanese is an important language, it has received a good deal of attention from scholars and other parties interested in its maintenance, and its intrinsic value within the larger Indonesian community.

It is possible to find articles pointing out its value, and to trends that indicate a potential decline in its use, negative language attitudes among younger people, and the way Javanese teaching is conducted in upper primary education in Java. Criticisms here have noted the less than ideal quality of Javanese language textbooks for Primary school children, the inadequate training of Javanese language teachers, the lack of adequate reference works such as dictionaries and grammars and the fact that only one hour per week of a total of 36 is usually allocated for Javanese language teaching.

Some organizations have created corpora of Javanese which can be studied on computer. This work is laudable and represents a major effort to record, transcribe and document the use of Javanese. Such Javanese corpora can probably be described as specialized rather than general, for example focusing on language acquisition in young children or specific regional or other varieties. However, to my knowledge, the idea that a concerted effort be made to create a large, comprehensive corpus of modern spoken and written Javanese that can be considered as broadly representative of the totality of Javanese with its geographical and social variation has received relatively little attention in the context of Javanese language rights and maintenance. In particular, it would be of great interest to understand the current state of both spoken and written Javanese that captured the important kinds of variation within it.

Corpus linguistics, an emerging set of methods in linguistics, is now mainstream and the development of corpora for language study is now quite advanced. Corpora are of use in studying various aspects of language such as the lexicon, semantics, morphology, phonology, syntax, and discourse. Corpora are also now provide the main source of data for the creation of modern dictionaries, reference grammars, and language teaching textbooks. Corpora are also essential for computational linguists who are developing language capable software such as web search engines, automatic summarization software, and translation software. Because of the link between the perceived need to produce a new generation of dictionaries, grammars and teaching texts of the contemporary state of language, and the proven value of corpora in doing just that, it should be obvious to Javanese scholars that corpus development for Javanese is something that deserves their attention.

The situation of minority indigenous languages in the UK is described in Lauder (2010a). The issue of language loss and maintenance in Indonesia is found in Lauder (2008b). The use of corpora for the study and maintenance of languages in Indonesia is covered in Lauder (2008a, 2010b, 2004a, 2008c) and Lauder and Lauder (2004).

This paper is therefore intended to introduce to an audience of expert scholars of Javanese about corpus linguistics and the development of corpora, what a corpus is, what types there are, what principles govern their design and construction and what the value of such work would be, in particular as related to the case of Javanese which is undergoing language change.

Corpus Linguistics

Corpus linguistics emerged in the 1960s and now, riding on advances in computer technology, corpus methods in the study of language have now attained mainstream status. Corpus linguistics is not a new branch of linguistics such as sociolinguistics or psycholinguistics. It is rather a broad set of tools, principles and research techniques based on the study of large quantities of text on computer. Corpus methods can be considered to be empirical and they can be used to investigate many types of linguistic questions. The term 'corpus' has a formal status and is used exclusively to mean an electronic corpus.

Language Corpora

There is no single definition of corpus but we can provide a stipulative definition for our present purpose as follows. A corpus[1] is a collection of naturally occurring[2] texts[3], stored in electronic format on a computer. The texts in the corpus are usually[4] assembled purposefully to form a representative[5] or balanced[6] sample of a language variety[7] or state. Corpora are used as research data to draw conclusions about that variety. Corpora are investigated using special software that can reveal information about words, such as their frequency, collocational patterning and statistical significance. See, for example, (Baker et al., 2006, Centre of Computational Linguistics, 2006, McEnery and Wilson, 2001, Sinclair, 1991).

The Uses and Applications of Corpora

Corpora are a useful form of data. They consist of large quantities of socially contextualized data that allow empirical analyses. Corpora can be of considerable use in a wide variety of theoretical, applied and interdisciplinary areas such as all branches of general linguistics, first and second language education, language maintenance, language policy and planning, cultural and gender studies, media and communication studies, computational linguistics and language engineering.

Corpora are of use in such applied linguistic areas as the creation of dictionaries, grammars, and language teaching materials. They can be used for the linguistic study of spelling, vocabulary, morphology, syntax and discourse. Through this, new grammars of the spoken language can be developed. They can be used in various studies of the lexicon, providing evidence of semantic patterning, word structure, and loan processes with other languages. The findings from corpus linguistics can be of great use in formulating language policy, and in language planning for expanding the lexicon and developing terminology. Corpora also play a central role in the creation of software which can perform language related tasks such as translation, web searches, text summarizing and so on.

Most of these applications or uses would apply to a corpus of Javanese. Of particular use for the current status of the language, and issues related to it, would be a Javanese corpus for the creation of dictionaries, grammars, and teaching materials.

Corpus Type and Purpose

A number of different kinds of corpus can be distinguished either according to their internal properties or their intended purpose(s). Commonly recognized types of monolingual corpora are GENERAL, SPECIALIZED, LEARNER, DIACHRONIC and MONITOR. However, these types do not represent mutually exclusive categories. Rather, we sometimes find that pairs of corpus types, such as general and specialized, represent different ends of a spectrum, where internal properties are shaped to varying degrees by slightly different purposes.

For example, an important factor with general and specialized corpora is the purpose or purposes for which they are intended. A GENERAL CORPUS is designed to be used for a wide variety of uses, such as language planning, linguistics, stylistics, applied linguistics, language teaching, lexicography, natural language processing (NLP). The general corpus therefore has to be broad in scope. One of

1 corpus (plural corpora) from Latin for 'body'.
2 naturally occurring, or attested in use in actual communicative situations, as opposed to being created specifically for inclusion in the corpus.
3 texts, here used in the sense of both written and spoken texts.
4 usually: this has been the case till recently but with the introduction of corpora created from the internet, this requirement has proved problematical and some linguists have omitted this as a defining feature of corpora.
5 representative, here in the sense commonly found in scientific work of a representative sample of a larger population.
6 balanced, meaning that the component parts are based on decisions about the criteria used for selecting texts and their proportions.
7 language variety, or varieties.

the main distinctions is whether it should contain only written language or also contain spoken language. It also means that it should contain a wide variety of GENRES[8], such as conversation, discussions, interviews, debates, speeches, essays, social letters, academic writing, popular writing, news reports, instructions, editorials and so on. The proportions of these different categories may also be balanced according to other categories, for example DOMAIN or REGISTER, in order to represent as wide a range of the language as possible. General corpora also will most likely use internationally agreed standards for encoding. An example of a general corpus is the British National Corpus (BNC) (McEnery et al., 2006: 59-60).

Meanwhile, a SPECIALIZED CORPUS is designed with a more specific purpose than a general corpus. It is intended to represent a specific variety of language, genre or domain which will be the object of study of the particular research endeavour. Examples of specialized corpora are for written English from the petroleum domain, or computer science. Another is The Michigan Corpus of Academic Spoken English (MICASE), a corpus of spoken academic English[9] (Simpson et al., 2002). Specialized corpora vary in size and composition according to their purpose(s). They can be relatively small, as in the case of a corpus of the works of one author, for example Shakespeare, or for frequency-based studies of grammatical behaviour. They may also be larger, for example to study particular specialist genres of language such as child language, or the language used by learners of English (McEnery et al., 2006: 60-61).

A general corpus may also be seen as a "standard reference" for the language variety which it represents and may thus be referred to as a REFERENCE CORPUS. This is because it is composed on the basis of relevant parameters agreed upon by the linguistic community and will usually include spoken and written, formal and informal language representing various social and situational strata. These corpora provide "a yardstick" for comparing successive studies with and as "a benchmark" for lexicons and for the performance of generic tools and specific language technology applications (McEnery and Wilson, 2001: 32).

Another dimension that underlies corpus design is the time period during which the texts were produced. The time frame the corpus represents can vary along two dimensions. The first is the length of time covered. This may be a relatively short period, with the corpus being seen as representing the language from a single short 'slice' of time. It may cover a larger 'chunk' (or chunks) of time. It may also stretch over very long periods of time and allow the study of language change. A DIACHRONIC (or historical) CORPUS contains examples of language from different time periods. It is used to allow tracking of language change over decades or centuries (McEnery et al., 2006: 65). Meanwhile, the term SYNCHRONIC CORPORA is used for corpora which are designed to represent national or regional varieties of the language, for example regional variants of Bahasa Indonesia, and which can be used to compare varieties, for example the Indonesian used in central Java as opposed to that used in Lampung. There are few synchronic corpora which allow the comparison of geographical variation (dialects) (McEnery et al., 2006: 64).

The next time-related dimension is whether the corpus is STATIC or DYNAMIC (changing). Corpora which are designed to represent the language from fixed period of time and which, once completed, no new material is added to, are thought of as "static". Many corpora are like this. Because the time frame it represents is fixed, the corpus can be designed to achieve a balance among its components. For example, in the BNC, the designers made conscious decisions about the proportion of texts to be included from different domains (imaginative, arts, belief and thought, commerce/finance, leisure and so on) and medium (books, periodicals, miscellaneous published, miscellaneous unpublished, and written to be spoken). Static corpora are useful as a reference for studying a particular time period of a variety or a language. The problem is, languages change over time.

---

[8] For clarification of the meaning of terms such as genre, register, text type, domain, sublanguage, and style as categories in corpus construction, see Lee, David. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language, Learning & Technology*, 5(3), 37-72.

[9] MICASE is a collection of nearly 1.8 million words of transcribed speech (almost 200 hours of recordings) from the University of Michigan. It contains data from a wide range of speech events (including lectures, classroom discussions, lab sections, seminars, and advising sessions) and locations across the university.

Eventually, a static corpus will no longer reflect the contemporary state of the language and will become out of date. If the corpus is intended to be relevant to the contemporary state of the language, then a static corpus will have a limited time frame of use. There is a type of corpus, however, which can keep track of the present state of the language because material is constantly or regularly being added to it annually, monthly, or daily (McEnery et al., 2006: 67). This is a MONITOR CORPUS. The monitor corpus was conceptualized and developed by Sinclair (1991: 24-26). Its key feature is that it is "dynamic", with the potential to keep increasing in size. This helps keep it current but makes it difficult to keep it balanced. The goal of achieving balance is to some extent achieved by sheer size.

In corpus design, many factors come into play and each may affect the others. A clear picture of all foreseeable purposes is essential before deciding on what the content will be and how it will be selected. In terms of a corpus which would be constructed to serve the purpose of representing the state of contemporary Javanese, either a static general or a monitor corpus seem to be the most appropriate kind to develop. However, it is also necessary to have some idea about other design criteria in relation to the state of Javanese, which is subject to both geographical and social variation. If the corpus was to be maximally useful for language planning and a range of other purposes, then it needs to consider in particular the issues of size, representativeness and balance, and that would necessarily involve a discussion about the nature of variation which the corpus must represent.

B. Corpus Design and Construction

A number of factors related to design need to be taken into consideration when planning such a corpus. In this section we look at corpus size, and that at the factors influencing the choice of content. What to include would be decided primarily around the issues of representativeness and balance as they relate to variation in and standardization of Indonesian.

Corpus Size

In the design of corpora, size is an important issue. Discussion of how large corpora should be is found in a number of publications (Atkins and Rundell, 2008, Biber et al., 1998, Kennedy, 1998, Krishnamurthy, 2002, McEnery et al., 2006, Ooi, 1998).

The size of a corpus is normally given in terms of the total number of words (TOKENS[10]) in it. During the 1960s and 1970s, corpus size was constrained to about a million words by practical considerations and this led to problems such as "data sparseness" (Atkins and Rundell, 2008: 57). Today, technical constraints no longer place a limit on corpus size which means that many of the criticisms leveled at the use of corpora in the 60s and 70s, for example by Chomsky, no longer are valid. Corpora have been increasing by a factor of one order of magnitude per decade since the 1970s. The Oxford English Corpus (OEC) broke the one billion[11] word barrier in the 2000s and is continuing to grow (Atkins and Rundell, 2008: 58). The question, 'What is the maximum size for a corpus?' doesn't have a definitive answer. On the one hand there are those who suggest that there is no upper limit on the size a corpus can be. Sinclair (1991: 18), for example, states that "A corpus should be as large as possible and keep on growing." However, there are also those who consider that increasing the size of corpora will not necessarily lead to better results for some types of research. As Leech (1991: 8-29) observes, "size is not all important."

It is more important to ask the question 'What is the minimum size needed?' This will depend on the purpose for which it is intended and also some practical considerations (McEnery et al., 2006: 71). Thus, the problem of size comes back to a matter of "descriptive adequacy" (Kennedy, 1998:

[10] token (corpus linguistics) a single occurrence of a word form in a corpus.

[11] billion: 1,000,000,000, a thousand million. This is the usage followed in the United States. In the UK, a billion is a million million, 1 followed by 12 zeros.

67). Corpus size depends on "the frequency and distribution of the linguistic features under consideration" for a particular purpose (McEnery et al., 2006: 72). Two distinct purposes are the study of lexis and the study of grammar. In general, the study of lexis require much larger corpora than the study of grammatical behaviour. The size of corpora required to perform quantitative studies of grammatical features can be relatively small because "the syntactic freezing point[12] is fairly low" (McEnery et al., 2006: 72). On the other hand, in contrastive lexical studies, to model the frequency distribution of a word, it is necessary to be able to contrast it with enough occurrences of others of the same category and this will require a much larger corpus. A corpus which was to be used in national language planning would necessarily be large because it would not be restricted to studies of syntax but also provide data on lexis, semantics, discourse and language variation.

The reason for the difference between required corpus size for grammatical and lexical studies lies in the regularities of the frequency distribution of words in language, something usually referred to as Zipf's Law. George K. Zipf[13] (1902-1950) was a Harvard professor of philology. In the 1930s, Zipf studied the word-frequencies of texts in English, German, Chinese and Latin. He noted "the orderliness of the distribution of words[14]" (Zipf, 1935) and found that "a few words occur with very high frequency while many words occur but rarely" (Zipf, 1935: 40), quoted in (Atkins and Rundell, 2008: 59).

In corpus linguistics, when counting the frequency of words in a corpus, it is necessary to distinguish between TOKENs, single instances of a graphic word (word form) occurring in a corpus and TYPEs, word forms, seen as distinct from other word forms. In a corpus with a million tokens there may be only around 25 thousand different TYPES (Hartmann and James, 1998). The term (word) TYPE is not specific about whether a word is to be treated separately or whether it is a member of a family of word tokens (Scott and Tribble, 2006: 13, Scott, 2007). Of all the TOKENS in a corpus, a small number of TYPES account for a large proportion of total tokens while a large number of types account for a very small proportion of all tokens. For example, in a corpus of 100 million words, with approximately 160,000 types, 8,000 types will likely occur 1,000 times or more each and account for 95% of all tokens in the corpus. Meanwhile, the remaining 152,000 types will only account for 5% of all tokens (Kennedy, 1998: 68).

Scott and Tribble (2006: 23) state that the 100 or 200 most frequent words in a corpus word list are mostly closed-set words, prepositions, determiners, pronouns, conjunctions. Medium frequency items come from frequency levels of around 5,000, 4,000 or 3,000 per 100 million words. At these frequencies, the words are all content words, nouns, verbs, adjectives (Scott and Tribble, 2006: 25). Meanwhile, 40% of the items in the frequency list will be HAPAX LEGOMENA, appearing only once (Scott and Tribble, 2006: 26). See also Kennedy (1998: 67) and Scott and Tribble (2006: 27, 29) who give the figure of around 50%.

Based on the general observation that the less the frequency, the greater the number of words, Zipf formalized this relationship in a mathematical formula (Oakes, 1998: 54-55, Pustet, 2004: 8, Scott and Tribble, 2006: 26, Zipf, 1965: 24). Zipf's Law, as it has become known, shows that there is a constant relationship between the ordinal rank[15] of a word in a frequency list, and the frequency

---

12 syntactic freezing point: the point at which any further increase in the size of a corpus would not produce new data that would be required for describing a particular syntactic feature.

13 Zipf: prounounced /zif/. The p is silent.

14 words: Zipf turns to the question of defining the term 'word' on page 39ff. He notes the ambiguity of the term 'word' in that *child* and *children* may be seen as either one word or two, and that *give*, *gives*, *given* as one or three. This is the difference between WORD FORM in the sense of a single member of a word family or "an inflectional variant of a lexeme" (See Jackson, Howard. 2002. *Lexicography: An Introduction*. London & New York: Routledge.) and LEMMA, which is a collection of systematically related word forms that are thought to share the same meaning (See Sinclair, J. M. 1991. *Corpus, Concordance, Collocation*. Describing English Language Series. ed. John Sinclair and Ronald Carter. Oxford: Oxford University Press.). It should be noted that Zip's terminology differs from that used today. He defines his use of terms on p. 40. He refers there to word forms as 'words' and to lemmas as 'lexical items'.

15 ordinal rank: this is the number for the sequence in a list based on the frequency of occurrence of the items. For example, if the words have the frequencies: 100, 50, 25, 12, 6, 3, then the rank order is 1, 2, 3, 4, 5, 6. It is possible that two or more items will have the same frequency.

with which it is used in a text. When all of the words (tokens) in a corpus are placed in rank order by descending frequency, and each rank is given a number, then the rank number (r), multiplied by the frequency (f) for each token will be approximately constant (C). This is expressed as ($r \times f = C$). The relationship, one of inverse proportionality, holds for most words except those of the highest and lowest frequencies (Crystal, 1997: 87).

Zipf's findings still hold good today. This means that in any corpus, about 40 percent of all of the word types will occur only once. Such single occurrences (HAPAX LEGOMENA) are not considered by lexicographers adequate for describing any word's behaviour. This means that in any corpus, there will not be enough data for about 40 percent of the words (types) to create an entry. Only those words (types) for which there is enough data will be able to find their way into the dictionary. Two questions therefore arise. The first is, 'How many times does a word (type) need to occur for it be considered as an adequate basis for description?' (Kennedy, 1998: 67), and 'How many words (tokens) would a corpus need to have for it to be big enough to supply enough words (types) for a dictionary with a specified number of headwords?' Krishnamurthy (2002), has set out a way of working out how large a corpus would be needed to produce different size classes of dictionaries, e.g. pocket, collegiate, unabridged and English for Foreign Learners (EFL). Krishnamurthy (2002) concludes that a corpus of 100 million words, the size of the British National Corpus, would be good enough for producing a Pocket Dictionary, but would "struggle to meet Collegiate requirements". He concludes that if one wishes to produce an Unabridged Dictionary, then a billion word corpus would be an entry level requirement, and that bigger would be better.

However, a billion word corpus may not be enough. This is because the calculations used by Krishnamurthy (2002) are based on the assumption that ten occurrences (TOKENS) of any lexical item would be adequate. But ten must be seen as a bare minimum. Firstly, high frequency words tend to be polysemous. The verb break occurs nearly 19,000 times in the BNC, but it has at twenty distinct senses can be distinguished for break along with a dozen phrasal verbs, some of which are also polysemous (Atkins and Rundell, 2008: 60-61). The word is also found in combination in many collocations, phrases, and grammatical patterns. Among these different phrases, patterns, and collocations, some are frequent and some rare. This means that some uses of break are important in a particular domain but rare in a large general corpus. For example, there are only eight occurrences of the phrase 'break someone's serve/service' (in the field of tennis) (Atkins and Rundell, 2008: 61). In my own experience of collocation analysis, between one and five thousand occurrences of a word (TOKEN) might be needed to obtain a hundred or so different collocates. It therefore appears simplistic to set a minimum requirement for a word's frequency in a corpus at ten or a hundred even five hundred. We need to take into account whether the word is polysemous, is involved in complex lexico-grammatical patterning, or in a rich phraseology.

If we take into consideration the continued growth trend of corpora, we can conclude that Krishnamurthy's (2002) figure of a billion words being suitable for producing an Unabridged dictionary probably no longer applies. Rather, a ten billion word corpus would be a more reliable starting point for such a dictionary and that even a corpus of this size should be seen as a starting point, as something that could be added to. However, a corpus of even ten billion words would not be much better than one of a billion if it consisted of only one type of text, for example newspaper articles. A general or monitor corpus needs to be more than large; its content must be representative.

Corpus Content

This section deals with the factors that are considered important when deciding what a corpus should contain.

Representativeness and sampling

The general or reference corpus has to provide data from which generalizations can be drawn about some variety of language or other. When the corpus is designed to study very large varieties or

"language" as a whole, then it would either have to contain everything or be a sample of that totality. Language, however, in the sense of a form of communication used by a community of speakers, would have to mean literally everything, all of the communicative events occurring, the totality of texts produced, spoken or written.

Any attempt to create a corpus of the Indonesian language would obviously be a huge and practically impossible task. If we wish to study large varieties or even 'language' itself, then sampling is the only option (McEnery and Wilson, 2001: 29). We must therefore attempt to make the corpus as representative of that totality as possible. Representativeness has been recognized as a fundamental issue in corpus construction (Barnbrook, 1996: 24, Biber, 1993: 243, Teubert and Cermáková, 2004: 113). The sample should be as representative as possible of the variety under investigation (McEnery and Wilson, 2001: 30) so that generalizations can be made about the whole population from the sample (Kennedy, 1998: 62, Manning and Schütze, 1999: 119).

In the social sciences, the population can usually be well-defined and is limited in extent. However, natural languages do not lend themselves to analysis using sampling and there are problems with using this approach when studying language. In order to obtain a representative sample from a population, it is necessary to define the population and the sampling unit (McEnery et al., 2006: 19). A sampling unit can be any unit of language, a book, a periodical, an article, a newspaper. The complete listing of these categories is referred to as the sampling frame (McEnery et al., 2006: 19).

Two approaches to creating sampling frames have been taken for building corpora of written texts. The first is based on a comprehensive bibliographical index. The sampling units for the LOB corpus was written British English text published in the UK in 1961. The sampling frame was taken from the British National Bibliography Cumulated Subject Index 1960-1964 for books and Willing's Press Guide for periodicals (McEnery et al., 2006: 19). The second is to define the sampling frame as the contents of a particular library which belong to the variety and period in which the researcher is interested in. For the above case, this might be defined as all the German-language books and periodicals in the Lancaster University Library which were published in 1993. This approach was taken for the Brown corpus (McEnery and Wilson, 2001: 78-79).

However, such catalogues and indexes do not exist for spoken language or non typed private correspondence. These are not published and so there would be no evidence of them in libraries. That means that this approach would not be helpful if the objective was to create a corpus of the spoken language. Further, the approach is likely to be of limited use in Indonesia because similar catalogues or guides are not available, and because the libraries do not hold rich, representative collections of texts. Clearly, for the construction of a large, general corpus of Javanese, some other sampling frame must be found.

Genres, registers, text types, domains and styles

Another approach in the attempt to create a representative sample of a language is to base the corpus around categories used in linguistics such as GENRE, REGISTER, TEXT TYPE, DOMAIN or STYLE. Take genre for example. In the British National Corpus (BNC), we find texts classified as face-to-face conversations, phone calls, classroom lessons, broadcast discussions, parliamentary debates, all of these being spoken, and student essays, social letters, press news reports, and novels, which are all written texts. The problem is that categories like genre are complex and linguists vary considerably in their definitions of them. An extremely good discussion of the nuances in the varied uses of the terms such as genre or register is given by Lee (2001). He considers text type to be an "elusive concept which cannot yet be established in terms of linguistic features" (Lee, 2001: 41). Genre and register he treats as "two different points of view covering the same ground" (Lee, 2001: 46) with genre looking at "memberships of culturally-recognizable categories" and registers as arising from "lexico-grammatical and discoursal-semantic patterns associated with situations". Although the two terms are sometimes used interchangeably, he prefers genre for classifying texts in corpora. Despite the fact that genres are usually recognizable as text categories, problems for classification are very real. One of the issues is that genres may come in different degrees of detail, something akin to the semantic taxonomies where we organize basic level terms with

superordinates and subordinates. For example, newspapers represent a genre of text, but contain multiple sub-genres, like news article, leader, editorial, or opinion piece. Further, novel might be seen as a basic level term, with literature as a superordinate category and western, romance novel and so on as subordinates (Lee, 2001: 48). Lee (2001: 49) proposes that genres in corpora are treated as basic level categories which can be characterized by seven attributes: domain (e.g. art, science), medium (e.g. spoken, written), content (topics, themes), form (e.g., generic superstructures), function (e.g., informative, persuasive), type (rhetorical categories: narrative, expository), and language (linguistic characteristics). He also mentions the possibility of adding: setting or activity type, and audience level. The discussion about terms such as genre, register and style continue to attract attention (Biber and Conrad, 2009).

Defining the categories and subcategories is certainly problematical. Genres are very complex, defying attempts to create neat taxonomies. Spoken genres tend to be harder to classify than written ones. The problem becomes even more acute when attempting to do the categorization by computer. This is because, while genre identification in linguistics may involve judgments about text-external factors, such as communicative purpose, computer identification must rely on text-internal, linguistic features. See for example Gunnarsson (2006) and Webber (2009). In addition to this, there are also problems in deciding how to assign texts to the different categories, or how to deal with texts that can be classified in overlapping categories (Atkins and Rundell, 2008: 64).

It might seem that one way around the problem of the definition of criteria used as a sampling frame for corpus construction was so problematical that it could be avoided altogether if corpora were constructed not on the basis of such socially determined categories but on the basis of the internal, or linguistic characteristics of the texts (Otlogetswe, 2004). However, a number of authors maintain that the text attributes or parameters used to do this should be extra-linguistic (external) and independent of linguistic criteria (Atkins et al., 1992: 5-6, Biber, 1993: 256, McEnery et al., 2006: 14, Sinclair and Ball, 1995). The reason for this, pointed out by Sinclair, is that any conclusions about text types based on word frequency distribution in such a corpus would be circular and invalid.

Considerations such as these have been discussed in great detail since the 1990s. Recently, the problem has been compounded by the emergence of new text types that are the result of the growth of new media such as the internet, social networking websites like FaceBook and smart phones with short messaging and chatting. Some of these new forms of communication don't fit neatly into the traditional written versus spoken mode distinction (Atkins and Rundell, 2008: 65). It turns out that it is practically impossible to define the whole population which the corpus is supposed to represent, and consequently, it would therefore be "logically impossible to establish what the 'correct' proportions of each component" should be (Atkins and Rundell, 2008: 66).

Variation in Javanese and the criteria for the sampling frame

Another factor that needs to be considered when sampling Javanese is that it exhibits both geographic and social variation. Linguistic variation is of importance in the design and analysis of corpora (Nevalainen and Taavitsainen, 2008).

The first kind of variation that needs to be considered is between the spoken and written forms. We next need to consider how to treat the regional varieties of Javanese. When the BNC was made, it sampled everyday, spoken conversational English from over thirty cities. Javanese is used primarily, but not exclusively on the island of Java, so decisions would have to be taken about where geographically to obtain samples.

The BNC also obtained spoken language data sampled according to socially situated categories such as business, academic, legal and so on. This is particularly important to capture the different semantic uses of words. For example, in English, the word sport is primarily used as a noun but also has a verbal form. An example of a default use of the verb is 'to sport a beard' or 'sport a new coat' where it carries the sense of 'have', or 'wear'. However, in other contexts, the connotation of

'is proud of' comes out. You can't capture such nuances of meaning if your data sample does not include the necessary domain or register.

In particular, it would be of importance in the case of Javanese to make a decision about what social levels are to be distinguished and how they would be identified.

Balance

While the goal of building a representative corpus is "unachievable", it remains a worthy "aspiration". Some principles can guide the process of corpus construction. One such principle would be to avoid sampling from only one text type, a "monolithic" corpus such as one constructed solely from news media texts. It is tempting to use news media corpora. They are easily available and often extremely large because of the availability of news in electronic form. However, this is to be avoided for general corpora because no matter how big a corpus of news texts is, it will never contain the kind of lexis that would be found in other genres, such as literary or academic. This means that it is advisable to sample from as wide a range of text types as possible. However, practical considerations constrain our ability to define and sample from all known text types to create a truly "representative" corpus. A modest, practical compromise between these two extremes would be to try to create a "balanced" corpus (Atkins and Rundell, 2008: 66).

A "balanced" corpus is a rational compromise but because it involves many subjective decisions and is also shaped by practical considerations like budgets and time frames, it can never be considered the result of a "scientific process" (Atkins and Rundell, 2008: 66). A "balanced" corpus, however, has a number of advantages. The use of good criteria allows the set up of a useful typology of text types. If stratified sampling is used to identify candidate texts for each text category, the result will be systematic and will reflect the actual types. If each text is labeled with information about its key features, such as genre, authorship, date of publication and so on, then users will be able to query subsets of the corpus to research how these things influence language use (Atkins and Rundell, 2008: 66).

Balance, the range of text categories in a corpus is a significant factor in representativeness, and what would be an "acceptable balance" is "determined by its intended uses" (McEnery et al., 2006: 16). A general corpus should contain a wide range of text categories which need to be sampled proportionally to some rational and explicit estimate of the population so that "it offers a manageably small scale model of the linguistic material which the corpus builders wish to study" (Atkins et al., 1992: 6). However, achieving balance is more "an act of faith" than a statement of fact or the result of some scientific measure. While there is no overall agreement on how to achieve balance, work in text typology, the classification and characterizing of text categories, is relevant to such attempts.

Another acceptable approach is to emulate existing corpora which are generally acknowledged to be balanced. The designers of the British National Corpus (BNC) have made a number of subjective decisions about balance but they have done so in a way that is as reasonable as any (McEnery et al., 2006: 17).

The BNC contains approximately 100 million words, of which 90 percent is from written sources and 10 percent from transcripts of speech. The texts in the written section are selected according to a sampling frame which considers domain, date and medium (e.g. book, periodical, etc). The spoken section has two parts: transcripts sampled demographically, and transcripts from context-governed situations. The demographic part consists of everyday conversation around the country whereas the context-governed section samples language in more formal settings such as education, business and institutional. The proportions of these components are shown in Table 1 and those for the spoken component in Table 2.

Table 1 Composition of the written BNC

Domain	%	Date	%	Medium	%
Imaginative	21.91	1960-74	2.26	Book	58.58
Arts	8.08	1975-93	89.23	Periodical	31.08
Belief and thought	3.40	Unclassified	8.49	Misc. published	4.38
Commerce/finance	7.93			Misc. unpublished	4.00
Leisure	11.13			To be spoken	1.52
Natural/pure science	4.18			Unclassified	0.40
Applied science	8.21				
Social science	14.80				
World affairs	18.39				
Unclassified	1.93				

Table 2 Composition In the Spoken BNC

Region	%	Interaction type	%	Context-governed	%
South	45.61	Monologue	18.64	Educational/informative	20.56
Midlands	23.33	Dialogue	74.87	Business	21.47
North	25.43	Unclassified	6.48	Institutional	21.86
Unclassified	5.61			Leisure	23.71
				Unclassified	12.38

Another corpus we can look to as an example for a future general corpus of Javanese is the privately created Standard Indonesian Language Corpus[16] (SILC). The corpus contains just over 26 million words of written contemporary Indonesian, making it suitable for exploratory work or for preliminary word list creation. The majority of the texts in the corpus, (79.21%), were published between 1996 and 2000 with 16.41% published between 1991 and 1995, 0.87% published between 1986 and 1990 and the remainder unclassified. Table 3 shows the composition of SILC. Texts are found in 8 broad domains which contain a total of 248 subject areas.

Table 3 Composition of the Standard Indonesian Language Corpus (SILC)

Domain	%
Social sciences and humanities	33%
Non-fiction (popular) and current events (news)	18%
Natural, applied and medical sciences	16%
Business (management, finance, economics)	10%
Literature (fiction, novels, poetry, drama)	10%
Religion and philosophy	8%
Lifestyle	3%
Art (fine arts, performing arts, music, decorative arts)	3%

In addition to being balanced by domain, the texts are all classified along a number of other dimensions. These include the following:

[16] SILC was created by the author, Multamia RMT Lauder and Muhadjir.

	Category	Parameters
1	Age (intended audience)	adult, teens, pre-teens, child
2	Audience (level)	academic, popular
3	Formality (treatment of work)	formal, neutral, informal
4	Publication (text type, genre)	book, conversation, drama, interview, etc[17]
5	Mode	speech, writing

The degree of variation in Javanese poses a challenge to corpus designers. If a balanced corpus was the goal, a good deal of consideration should be given to looking at the available information about what criteria should be used as the basis of the sampling frame. Take 'region', a category used in the BNC, and which had just three divisions: south, midlands and north. Would three regional varieties of Javanese be acceptable or would finer distinctions be made?

Even though achieving representativeness and balance may be highly problematical, this cannot be seen as a reason for avoiding corpus linguistic work or for dismissing the results of corpus analysis as unreliable or irrelevant (McEnery et al., 2006: 19). Corpora have been for some time and are likely to grow in usefulness in a number of areas of investigation.

Texts from the internet

The internet is increasingly being seen as a readily available source of extremely large quantities of text which are already in electronic format. For example, many newspapers and magazines have online versions of their print editions and despite the fact that the number of web pages in the Indonesian language is far fewer than those in English, the number is being added to all the time. It has been suggested that web-based corpora such as the Oxford English Corpus (OEC), while not balanced, have attained such a size that this no longer matters. Supporters of web-based corpora have produced some preliminary research indicating that web-sourced corpora compared favorably with benchmark collections such as the BNC (Fletcher, 2004). However, it is probably best to be cautious with this view when it comes to a Javanese corpus because the status of Javanese material online is probably not comparable to that for English. There may be a good number of Wikipedia articles in Javanese. However, it is likely, for example, that a number of domains, such as literature, are poorly represented. Further, we have to question whether a corpus of Javanese sources solely from the web could be representative of the population of users of Indonesian when only a small percentage of Javanese speakers are connected to the web and only a small number of these are posting content online. Despite the usefulness of these materials, to give too much prominence to these in the corpus seems to me to be wrong.

Copyright

It would be difficult to construct a large general corpus without using any copyright material at all. Using copyright material without permission could place corpus builders in the position of being liable for claims by the owners of loss of profits and can be

When corpus builders use copyright material without permission, they may be liable to being taken to court by the copyright holders who may claim that they have suffered a loss of profit because its use in the corpus reduces their ability to sell it (McEnery et al., 2006: 77). Corpora can be seen as having two distinct purposes: those designed purely for academic research, and those designed for commercial reasons, such as the creation of dictionaries or software. However, large corpora are expensive to make and they might be used for both reasons.

Copyright law is complicated and it also varies from country to country. The general advice is to get permission, even for non-profit corpora (McEnery et al., 2006: 78). Detailed suggestions of how to go about this are given in Atkins and Rundell (2008: 81-84). In Indonesia the most recent law is Law Number 19 of 2002 on Copyright. According to Kusumadara (2008), the Law has been

[17] Publication type or genre: anthology, book, booklet, conversation, debate, dictionary, discussion, drama, interview, journal, leaflet, magazine, manual, news magazine, newspaper, novel literary, novel popular, personal writing, poetry, press release, prospectus, reference work, report, textbook.

difficult to enforce because the concept of intellectual property is still foreign to the country and they are not appropriate to its current stage of development. The situation in Indonesia is more recently reviewed in (Antons, 2009). However, it is not clear how the 2002 Law applies to the use of published texts in corpora. The 1997 Law on copyright in Indonesia, Article 14 indicates that it is not an infringement of copyright to use an authored work for education or research, as long as it is cited and the "normal interest" of the author is not prejudiced.

C. Conclusion

The Javanese language is an important world language with a rich literary heritage and despite being an indigenous language in Indonesia, the term 'minority' hardly seems appropriate. However, despite the large number of speakers of Javanese, the language is not immune to the processes of language change.

Javanese is a linguistic and cultural treasure. It is one of the world's major languages. Yet its status as a regional language puts and its relationship with Indonesian are not helping it continue as a vibrant, relevant living language. In order to continue as the language which Javanese remain proud of, it must adapt to social changes and find ways to counter potential or emerging threats that could push it from its premier position into decline.

The forces that have a potential to do this include inadequate official support and a language policy that 'lacks teeth' in valuing the regional language and ensuring their protection. Also affected by policy, in the area of education, is inadequate space given to teaching the language, and problems with the provision of quality teaching materials or training of teachers. Attitudes of young people towards the language are important. If they see it in a negative light, they will be less motivated to use it and the chain of transmission could be broken. The relevance of the language is important. If Javanese were able to adapt for use in a wider set of domains and social situations, especially modern ones, then it would be more likely to win over the young. There may also be problems discouraging writers from producing creative works if there is poor enforcement of copyright protection. It is writers who are capable of taking the lexicon of Javanese and using it to create new Javanese words to express new meanings. This would be necessary to expand the lexicon to meet the challenge of communicating new ideas. Finally, in an era of globalization, it makes sense for the Javanese to go on line not only to search for information produced by others, but to put their own ideas out into the global ideas marketplace. To do this, it will be necessary to language software such as web search engines, translation tools and the like. These already exist for Javanese, but they require improvement. The experience of other indigenous languages shows that it is the community themselves who must take responsibility for the maintenance of their language. Local initiatives, both from local government and civil organizations can make a huge difference.

Linguists, more than anyone, are in a position to support the language. The creation of a large, general corpus of modern, contemporary Javanese could be of great use not only to its study, but also to answer questions about the nature and extent of the lexicon, the grammar of everyday spoken Javanese, the distribution of the use of different social levels, and other important linguistic questions. A corpus could form the basis for a number of practical applications relevant to language policy and planning issues, such as the creation of modern dictionaries, grammars, and language teaching texts.

The development of corpora is not a small undertaking and considerable resources and time are required. However, a Javanese corpus would be a treasure that is worthy of the language and the culture it embodies, and it would be much appreciated by academics, policy makers and writers.

References

Alwi, Hasan, and Sugono, Dendy. (2000). *Dari Politik Bahasa Nasional ke Kebijakan Bahasa Nasional* (From National Language Politics to National Language Policy). In Politik Bahasa : Risalah Seminar Politik Bahasa (Language Politics : Procedings of the Seminar on Language Politics). eds. Hasan Alwi and Dendy Sugono, v-xiv. Jakarta: Pusat Bahasa dan Departemen Pendidikan Nasional.

Antons, Christoph. (2009). *Copyright law reform and the information society in Indonesia.* In Intellectual Property in Asia: Law, Economics, History and Politics, Volume 9. eds. P. Goldstein, P. Ganea, J. Straus, T. V. Garde and A. I. Woolley, 87-128. Berlin & Heidelberg: Springer-Verlag.

Atkins, Sue, Clear, Jeremy, and Ostler, Nicholas. (1992). *Corpus design criteria.* Literary and Linguistic Computing, 7(1), 1-16.

Atkins, Sue, and Rundell, Michael. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Baker, Paul, Hardie, Andrew, and McEnery, Tony. (2006). *A Glossary of Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Barnbrook, Geoff. (1996). *Language and Computers : A Practical Introduction to the Computer Analysis of Language.* Edinburgh Textbooks in Empirical Linguistics. Edinburgh: Edinburgh University Press.

Biber, Douglas. (1993). *Representativeness in corpus design.* Literary and Linguistic Computing, 8(4), 243-257.

Biber, Douglas, Conrad, Susan, and Reppen, Randi. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge Approaches to Linguistics. ed. Jean Aitchison. Cambridge: Cambridge University Press.

Biber, Douglas, and Conrad, Susan. (2009). *Register, Genre, and Style.* Cambridge Textbooks in Linguistics Cambridge: Cambridge University Press.

Burch, Stella J. (2009). *Regional Minorities, Immigrants, and Migrants: The Reframing of Minority Language Rights in Europe.* Ms., Yale Law School Student Scholarship Papers.

Centre of Computational Linguistics. (2006). *Systematic Dictionary of Corpus Linguistics.* Kaunas, Lithuania: Centre of Computational Linguistics, Vytautas Magnus University. Online Address: http://donelaitis.vdu.lt/publikacijos/SDoCL.htm. Accessed.

Crystal, David. (1997). *The Cambridge Encyclopedia of Language, 2nd Edition.* Cambridge: Cambridge University Press.

Fletcher, W. H. (2004). *Making the Web more useful as a source for linguistic corpora.* In Applied Corpus Linguistics: A Multidimensional Perspective. eds. U. Conner and T. Upton, 191-205. Amsterdam: Rodopi.

Grimes, Barbara F. ed. (2000). *Ethnologue: Languages of the World.* Dallas, Tex.: Summer Institute of Linguistics.

Gunnarsson, Mikael. (2006). *Genre Identification on the Web, Academic Article: Swedish School of Library and Information Science,* Swedish National Graduate School of Language Technology.

Hartmann, R. R. K., and James, G. (1998). *Dictionary of Lexicography.* London: Routledge.

Jackson, Howard. (2002). *Lexicography: An Introduction.* London & New York: Routledge.

Kennedy, Graeme D. (1998). *An Introduction to Corpus Linguistics.* London and New York: Longman.

Krishnamurthy, R. (2002). *Corpus size for lexicography.* Corpora-List. Online Address: http://www.hit.uib.no/corpora/2002-3/0254.html. Accessed: 25 July 2010.

Kusumadara, Afifah. (2008). *Problems of enforcing intellectual property laws in Indonesia.* Paper read at The Law of International Business Transactions: A Global Perspective. Hamburg, Germany. April 10-12 2008. International Association of Law Schools (IALS), Washington, DC.

Lauder, Allan F. (2007). *Indigenous Languages in Indonesia: Diversity and Endangerment.* In Procedings of Kongres Linguistik Nasional XII. Surakarta, 3-6 September. Masyarakat Linguistik Indonesia & Universitas Sebelas Maret.

Lauder, Allan F. (2008a). *Principles of corpus design for local languages.* In Workshop on Empowering Local Languages through ICT. Bali, 26th August 2008. Kementrian KomInfo.

Lauder, Allan F. (2010a). *Language planning in a multilingual UK.* In Simposium Perencanaan Bahasa (SIPB) 2010: Perencanaan Bahasa pada Abad ke -21: Kendala dan Tantangan. Jakarta, 2—4 November 2010. Pusat Bahasa, Kementerian Pendidikan Nasional Indonesia.

Lauder, Allan F. (2010b). *Data for lexicography: The central role of the corpus [October 2010]. Wacana,* 12(2: Lexicology and semantics).

Lauder, Multamia R.M.T. (2004a). *Optimalisasi Bahasa Indonesia berbasis korpus linguistik*. In Pertemuan Ilmiah Bahasa dan Sastra Indonesia (PIBSI) XXVI.

Lauder, Multamia R.M.T, and Lauder, Allan F. (2004). *Upaya pembuatan korpus linguistik: Persoalan dan pemanfaatannya.* In Pertemuan Linguistik Bahasa dan Budaya Atmajaya (PELBBA) 17. Jakarta. Unika Atma Jaya.

Lauder, Multamia R.M.T. (2008b). *Preservation and maintenance of regional languages and scripts*. In Workshop on Empowering Local Languages through ICT. Bali, 26th August 2008. Kementrian KOMINFO.

Lauder, Multamia R.M.T. (2008c). *Preservasi dan pemberdayaan bahasa daerah*. In Seminar on Empowering Local Language Through ICT. Jakarta, 11 Agustus 2008. Departemen Komunikasi dan Informatika.

Lauder, Multamia RMT. (2004b). *Pelacakan Bahasa Minoritas dan Dinamika Multikultural* (Tracking the Minority Languages in a Multicultural Dynamic). In Simposium Internasional Kajian Bahasa, Sastra, dan Budaya Austronesia III (3rd International Symposium on Austronesian Language, Literature and Culture), . . Denpasar, Bali, 19-21 August 2004.

Lee, David. (2001). *Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle*. Language, Learning & Technology, 5(3), 37-72.

Leech, Geoffrey N. (1991). *The State of the Art in Corpus Linguistics.* In English Corpus Linguistics: Studies in Honor of Jan Svartvik eds. K. Aijmer and B. Altenberg, 8-29. London: Longman.

Lewis, M. Paul ed. (2009). *Ethnologue: Languages of the World.* Sixteenth edition. Dallas, Tex.: SIL International.

Manning, Christopher D., and Schütze, Hinrich. (1999). *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

McEnery, Tony, and Wilson, Andrew. (2001). *Corpus Linguistics: An Introduction, Second Edition.* Edinburgh Textbooks in Empirical Linguistics Series. ed. Tony McEnery and Andrew Wilson. Edinburgh: Edinburgh University Press.

McEnery, Tony, Xiao, Richard, and Tono, Yukio. (2006). *Corpus-based Language Studies: An Advanced Resource Book.* London and New York: Routledge.

Nevalainen, Terttu, and Taavitsainen, Irma eds. (2008). *The Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present.* vol. Volume 2. Studies in Language Variation: John Benjamins Pub.

Oakes, Michael P. (1998). *Statistics for Corpus Linguistics*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh: Edinburgh University Press.

Ooi, Vincent B. Y. (1998). *Computer Corpus Lexicography*. Edinburgh Textbooks in Empirical Linguistics Series. ed. Tony McEnery and Andrew Wilson. Edinburgh: Edinburgh University Press.

Otlogetswe, T. (2004). *The BNC Design as a Model for a Setswana Language Corpus.* In Proceedings of CLUK '04: 193-198. Birmigham, UK.

Pustet, R. (2004). Zipf and his Heirs. Language Sciences, 26(1), 1-25.

Renandya, Willy A. (2000). *Indonesia. In Language Policies and Language Education: The Impact in East Asian Countries in the Next Decade.* eds. Wah Kam Ho and Ruth Y. L. Wong, 113-137. Singapore: Times Academic Press.

Scott, Mike, and Tribble, Christopher. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education.* Studies in Corpus Linguistics, v. 22. ed. Elena Tonigni-Bonelli. Philadelphia: John Benjamins Publishers.

Scott, Mike. (2007). *WordSmith Tools 5.0 Help File.* Oxford: Oxford University Press.

Simpson, R. C., Briggs, S. L., Ovens, J., and Swales, J. M. (2002). *The Michigan Corpus of Academic Spoken English.* Ann Arbor, MI: The Regents of the University of Michigan.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation.* Describing English Language Series. ed. John Sinclair and Ronald Carter. Oxford: Oxford University Press.

Sinclair, J. M., and Ball, J. (1995). *Text Typology (External Criteria),* Draft Version. Pisa EAGLES ftp server, Birmingham. Online Address: Accessed.

Skutnabb-Kangas, Tove. (2000). *Linguistic Genocide in Education, or Worldwide Diversity and Human Rights?* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Teubert, Wolfgang, and Cermáková, Anna. (2004). *Directions in Corpus Linguistics.* In Lexicology and Corpus Linguistics: An Introduction, 113-165. London and New York: Continuum.

Wahab, Abdul. (2006). *Masa Depan Bahasa, Sastra, dan Aksara Daerah*. Jakarta: Pusat Bahasa. Online Address: http://pusatbahasa.diknas.go.id/laman/nawala.php. Accessed: 23 February 2011.

Webber, Bonnie. (2009). *Genre meets Text Technologies*, PowerPoint presentation: School of Informatics, University of Edinburgh.

Zipf, G. K. (1935). *The Psychobiology of language: An Introduction to Dynamic Philology.* Boston: Houghton Mifflin Company.

Zipf, G. K. (1965). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* (Facsimile of 1949 edition). New York: Hafner.