**ANNOTATION MODEL FOR LOANWORDS IN INDONESIAN CORPUS**

**: A LOCAL GRAMMAR FRAMEWORK**

Prihantoro

Universitas Diponegoro

prihantoro@undip.ac.id

*Abstract*

There is a considerable number for loanwords in Indonesian language as it has been, or even continuously, in contact with other languages. The contact takes place via different media; one of them is via machine readable medium. As the information in different languages can be obtained by a mouse click these days, the contact becomes more and more intense. This paper aims at proposing an annotation model and lexical resource for loanwords in Indonesian. The lexical resource is applied to a corpus by a corpus processing software called UNITEX. This software works under local grammar framework (Gross, 1993 & 1997). The lexical resource has already been tested on a small corpus to perform automatic retrieval of loanwords in Indonesian. The automatic retrieval aims at identifying the loanwords from different languages. The queries demonstrated in this paper allows the users not only to retrieve loanwords, but also show etymological information about the loanwords, mainly the donor language, and in some cases, the language that introduce the loanwords to Indonesian.

1. **INTRODUCTION**

A considerable number of words in Indonesian are borrowed from other languages. The donor language is not only limited to vernacular language, but also from languages overseas such as: *kursi* 'chair' and *terjemah* 'translate', which are adapted from Arabic words *kursi* and *tarjamah*. To some extent, loanwords show the degree of language contact.

Loanwords studies may vary, but one valuable study is the etymological research of loanwords. It aims at retrieving the source language and the transfer process. It is also aimed to investigate some problems related to the annotation of loanwords in Indonesian as well as proposing a model that can overcome the problems. The lexical resource which is used for annotation in this study is maintainable and open for enrichment. The result of this study is used to amplify the descriptive power of dictionary entries. Consider the entry *ajek* in Indonesian standard dictionary, which means to stay still. This entry is annotated with [Jw] code, suggesting that this lemma is borrowed from Javanese.

Unlike standard dictionary, which focuses on word entry level, the annotation model here aims on amplifying text analysis of a corpus. The aim is to create a lexical resource for loanwords in Indonesian, so that it can be applied to annotate a text corpus. The annotation will amplify the power of linguistic query in the Natural Language Processing (NLP) tasks such as opinion mining, information retrieval or extraction. The lexical resource is open for further development or application on different corpus. The lexical resource focused on simple loanwords both uninflected and inflected forms.

1. **LOANWORDS**
	1. **Etymological Research of Indonesian Loanwords**

Jones (2007) research on loanwords amplified the research to the process of borrowing itself. It is considered more detail as one can understand how a word travels from its donor language via different langauges. Consider *razia*, an entry in the dictionary of loanwords by Badudu (2003). This entry is listed under *Ar* code, which is the same like the finding of Jones. Jones, however, believed that even though the word comes from Arabic, *razia* is not a direct borrowing. Instead it was borowed by French, and introduced to Indonesian by Dutch.

1. **razia** [(police) raid] < Du *razzia* (< Fr < Ar) (var. Rasia)

Example (1) is the headword (entry) for the loanword *razia.* This entry shows that it usually collocates with *police* has another variation, which is *rasia*. This suggests that the word is brought by Dutch (Du), borrowed by French (Fr), but the donor language is Arabic (Ar), as it is shown by the rightmost code Ar. In Jones (2007) the entry for simplex and compound loanwords are marked. Consider example (2) to (5):

1. **dwifungsi** [dual function] < *dwi-* (< Skt) + < *fungsi* (< Du)
2. **widyaiswara** [teacher; civil service training officer] < *widya* (<Skt) + Skt *isvara*
3. **tunaaksara** [illiterate] < *tuna*- (< Skt) + < *aksara* (< Skt)
4. **tabula rasa** [tabula rasa] < Lat *tabula rasa*

Examples (2) to (5) are the examples for loanwords entry. The might look superficially similar but there are some notable differences. The lingual component marked by ‘-‘ indicates that they are bound as in *dwi-* and *tuna-*. The compounding component might be borrowed from the same language, for instance Sanskrit as shown by example (3) and (4). However, one of the components might be borrowed from different language as shown by (2), where *dwi-* is from Sanskrit and *fungsi* from Dutch. Unlike example (2), (3) and (4), where the blend is marked by ‘+’, example (5) is not marked by ‘+’. The reason is not that it is borrowed from the same langauge (*Latin*), but because they are ortographically separated by space: *tabula\_rasa*.

The result of Jones’ research is published by printed book as well as on-line resource, which makes it a kind of open access electronic dictionary. By having the dictionary in electronic form, the size and thickness (in other words, the portability) of a dictionary is no longer a problem (Chen, 2010). The indexing method also changes from alphabetical search to query based search. Its application version also allows the users to select which information to display that the users feel necessary or conformable with.

* 1. **Loanword and Text Analysis**

In the previous sub-section, we have discussed how the electronic dictionary can be a powerful resource for text analysis. The remaining question is how to use the electronic dictionary to support our text analysis. When the size of the text is small, the use of manual observation is of course more effective. However, when the size is improved, let us say 30 scripts, this will be a laborious work. For this reason, we need to process this text with the help of corpus processing software.

In this research, the corpus processing is performed by UNITEX (Paumier, 2008), a local grammar based corpus processing software. The Local Grammar approach by Gross (1993 & 1997) aims at describing the morphosyntactic behaviour, and to some extent semantic, by using finite automata. Local grammar based software has been successfully used in performing various Natural Language Processing (NLP) tasks in many different languages, for instances, the building of English lexical resource by Monceaux (1995), The recognition of Arabic proper names (Traboulsi, 2009), the automatic retrieval of archaic lexicon (Prihantoro, 2011), and the annotation of multiword expressions in French (Laporte, Nakamura, & Voyatzi, 2008). It was also used in researching corpus on languages with their own writing system like Korean (Nam, 2010).

To date, researching Indonesian corpora by using Local Grammar based corpus processing software is still rare. The auto-prefixing model for Indonesian (Prihantoro, 2011) is the first work that purely used local grammar based software. Verb annotation is one of the tasks that the software successfully performed. However, the annotation is purely grammatical, with no etymological information provided. As the number of loanwords is abundant, creating a lexical resource for annotation is a crucial importance because it might show the degree of language contact in Indonesian.

1. **METHODOLOGY**

A corpus might take the form of either raw or annotated. However, a corpus would have been more useful when it is annotated as the NLP tasks work beyond statistic and characters match, such as automatic retrieval, opinion mining, syntactic parsing and etc.

First and foremost, a lexical resource must be created. Building the lexical resource can be performed more conveniently with spreadsheet. However, it must be converted to txt extension to comply with UNITEX structure. The formalization of this lexical resource is designed to annotate loanwords both in canonical (root) and inflected form. The existing Indonesian lexical resource that complies with data structure used in UNITEX aims at annotating verbs with {meN-} prefix. This existing resource can be compiled to enrich the loanword lexical resource. When the lexical resource is applied, it is expected to annotate the corpus so that retrieval of loanwords can be performed.

The next step is testing the accuracy of the annotation by means of writing queries for automatic retrieval. There are two retrieval tools provided by UNITEX; regular expression and local grammar graph. Writing queries for regular expressions works relatively quicker, but it might not be sufficient when the constraints for the retrieval are complex. For instance, retrieval by using morphological mode cannot be performed by using regular expression. Therefore, queries in this paper are written for both. The result of the retrieval is shown by using concordance. To summarize this procedure, please refer to figure 1:

**Figure 1. Research Flowchart**

Raw

Corpus

Annotate

Results

By Concordance

Annotated

Corpus

Loanwords Lexical Resource

* Inflected Form, Ex: <karyawan>
* Uninflected Form, Ex: <karya>

Sending Queries

* Regular Expression
* LGGs
1. **ANNOTATION MODEL AND SYSTEM APPLICATION**

The lexical resource in this research excludes compound forms and focus on simple form resource. The simple form resource itself contains entries for both inflected and root form. Consider some of the entries illustrated by table 1:

**Table 1. Lexical Resource of Simple Form Loan Words**

|  |  |
| --- | --- |
| Entries  | Simple  |
| Root | <karya>, <panen>, <sila>, <komputer>, <fungsi> |
| Affix and Root | Root + suffix | Prefix + root |
|  | <karya> + {-wan} = karyawan | {pasca-} + <panen> = pascapanen{panca-} + <sila> = pancasila |

The entries presented by table x is still raw and unstructured. The entries do not explicitly show the sense or the origin. The roots on table x originate from different languages such as: English, Dutch and Sanskrit. Borrowing does not only take word, but also another morphologically smaller component, affix, such as {wan-}, {pasca-}, and {panca-}. However, this information is missing in table x. For this reason, the entries must be carefully structured equipped with useful information. In other words, the entries require annotation.

The annotation is performed to formalize the resource so that it is compatible with the corpus processing software used in this research, UNITEX, which work under local grammar framework. Please consider root (uninflected) form entries presented by figure 2:

**Figure 2. Uninflected Root Entry Lines**

kasino,NL0+CASINO+ITA+DU

sekolah,NL0+SCHOOL+POR

mutakhir,AL0+RECENT+AR

sila,NL2+PRINCIPLE+SKT

stop,VL0+STOP+DU

warta,NL1+NEWS+SKT

Figure 2 shows the annotation structure, where each entry line complies with UNITEX lexical resource format. The entry line begins with a lemma (headword), followed by a comma. The first uppercase character is the grammar code of the lemma. For instance, N stands for noun, V for stands for verbs and A for adjective. The second uppercase is an additional grammar code, but in this case, is the code for loanwords. Therefore, the grammar codes always end with L. What follows the grammar and loan code is the inflectional code (number) and its near-equivalence in English. Each code is separated by a plus [+]. The etymological code(s) lies after the meaning. For instance, the language code [+POR] for <sekolah> indicates that this root is borrowed from Portuguese [+POR].

When there are two language codes, the later is not the donor language, but the language that introduced the word to Indonesian. By referring to the entry line <kasino>, you will find [+ITA+DU] as the language codes. This informs us that the donor language is Italian [+ITA], but the word is introduced to Indonesian via Dutch [+DU]. As it has been commented previously, the borrowings might also take affix. Indonesian borrows a great deal of words, and some words are really productive by means of affixing. But some of these affixes are loans as well.

In table 1, the root like <karya> takes prefix, while some others like <panen> and <sila> take suffixes. On this ground, the inflection code must be distinguished. The grammar that code ends with [1] takes suffix, while [2] takes prefix. They are inflected automatically by inflectional LGGs. Consider figure 3:

**Figure 3. Inflectional LGGs for NL1 and NL2**



The inflection of these LGGs might include more semantic codes as some also enrich the semantic feature of the word. For instance, suffix {-wan} has the feature of masculine, while {-wati} is the feminine one. The semantic codes for affix are written in lowercase. Therefore when there are two semantic codes both in upper and lowercase, the lower case describes the property of the affix, while the upper case describes the property of the root. Please consider the inflected entry lines in (6) and (7):

1. wartawan,warta.NL+NEWS+SKT+male+skt:x
2. dwifungsi,fungsi.NL+FUNCTION+LAT+DU+two+skt:y

In (6), both the root and the affix are borrowed from Sanskrit. In (7), the root is from Latin, but introduced by Dutch. The affix is not borrowed from Latin or Dutch, but from Sanskrit. Here, the root and the affix originate from different language. When the corpus is already annotated, automatic retrieval is possible. For instance, by using this regular expression <NL>, the user can retrieve all noun loanwords in the corpus as presented by figure 4:

**Figure 4. The Result of Automatic Retrieval for Noun Loanwords**



Besides using regular expression, automatic retrieval can also be performed by using LGGs. The retrieval by using LGGs provides the users with some benefits. It allows users to retrieve with more complex constraints. One of the examples is morphological form retrieval. For instance, the retrieval seeks to find loan nouns, with loan affix. Consider the result presented by figure 5:

**Figure 5. The Result of Morphological Retrieval**



The result presented by figure x, is the retrieval for roots inflected by suffixes {-wi}, {-wan} and {-wati}. When displaying grammar and semantic codes is necessary, users can also specify the retrieval by using LGGs. The result shows all entry line codes. Please refer to figure 6:

**Figure 6. Entry Lines Codes Shown by Automatic Retrieval**



As it has been commented previously, the distinction between upper and lower case codes is really important as it may show the donor language from both the root and the affix. Consider the inflected form <dwifungsi> in figure x. It shows that <fungsi> originates from Latin, and the prefix {dwi-} is from Sanskrit.

1. **CONCLUSION**

This paper has shown a successful demonstration of how an Indonesian loanword lexical resource can be used to annotate a corpus. It managed to automatically retrieve loanwords, both the inflected and uninflected forms. This resource can be compiled to the previously existing Indonesian local grammar resource for Indonesian (Prihantoro, 2011). It is also open for enrichment both from the size and the structure. For the further study, I suggest that the annotation structure is expanded to comprehensively deals with compounds, as it has been commented previously on section 2. On this study, compound loan nouns can be retrieved by using LGGs in terms of their composition. In fact, the loanwords are not always compositional.

Another issue to resolve is the retrieval of loanwords that are 100% composed of bound morphemes. The lexical resource in this paper comprises of morphological resources for retrieving bound morphemes. However, the morpheme is always inflected to a free one. It is possible that a word is composed of all bound morphemes that are borrowed from different languages, such as <mikrolet> that are composed of {mikro-} and {-let}. Consider the meaning that is not semantically compositional, and unpredictable. False assumption can be made of {-let} to mean minute. However, it is a shortened from of <opelet> a kind of public transportation car. The lexical resource needs quantity enrichment as well constant improvement to deal with these challenges.

**REFERENCES**

Badudu, J.-S. (2003). Kamus Kata-Kata Serapan Asing Dalam Bahasa Indonesia. Jakarta: Penerbit Kompas.

Badudu, J.-S. (2008). Morfologi Bahasa Indonesia: Pendekatan Proses. Bandung: Rhinneka Cipta.

Chen, Y. (2010). Dictionary Use and EFL Learning. A Contrastive Study of Pocket Electronic Dictionaries and Paper Dictionaries. International Journal of Lexicography , 275-306.

Gross, M. (1993). Local Grammars and Their Representations by Finite State Automata. In M. Hoey, Data, Description, Discourse: Papers on the English Language in Honour of John Mc.Sinclair (pp. 26-38). London: Harper Collins.

Gross, M. (1997). The Construction of Local Grammars. In E. Roche, & Y. Schabes, Finite State Language Processing (pp. 329-354). Massachusetts: MIT Press.

Jones, R. (2007). Loan-Words in Indonesian and Malay. Leiden: KITLV Press.

Laporte, E., Nakamura, T., & Voyatzi, S. (2008). A French Annotated Corpus for Multiword Nouns. Language Resources and Evaluation Conference, (pp. 27-31). Marrakech.

Monceaux, A. (1995). Le dictionnaire des mots simples anglais : mots nouveaux et variantes orthographiques, rapport technique IGM 95-15. Paris: Institut Gaspard Monge.

Nam, J.-S. (2010). Application of Finite Graphs LGG in Information Extraction. Colloque organisé par l’Observatoire de linguistique Sens-Texte (pp. 49-59). Montreal: Université de Montréal.

Paumier, S. (2008). Unitex Manual. Paris: Universite Paris Est Marne La Valee & LADL.

Prihantoro. (2011). Local Grammar Based Auto Prefixing Model for Automatic Extraction in Indonesian Corpus (Focus on Prefix MeN-). Proceedings of International Congress of Indonesian Linguists Society (KIMLI) (pp. 32-38). Bandung: Universitas Pendidikan Indonesia Press.

Prihantoro. (2011). Transducer for Auto-Convert of Archaic to Present Day English: A Support for Computer Assisted Language Learning. Proceeding for The 16th English in Southeast Asia Conference. Yogyakarta: Sanata Dharma University Press..

Traboulsi, H. (2009). Arabic Named Entity Extraction: A Local Grammar Based Approach. Proceedings of the International Multiconference Science and Information Technology (pp. 139-143). Mragowo, Poland: Polish Information Processing Society.