# TRANSDUCER FOR AUTO-CONVERT OF ARCHAIC TO PRESENT DAY ENGLISH FOR MACHINE READABLE TEXT: A SUPPORT FOR COMPUTER ASSISTED LANGUAGE LEARNING[i]

PRIHANTORO
Universitas Diponegoro, Indonesia

**Abstract**

There exist some English literary works where some archaic words are still used; they are relatively distinct from Present Day English (PDE). We might observe some archaic words that have undergone regular changing patterns: for instances, archaic modal verbs like *mightst, darest, wouldst*. The *–st* ending historically disappears, resulting on *might, dare* and *would*. (*wouldst* > *would*). However, some archaic words undergo distinct processes, resulting on unpredictable pattern; The occurrence frequency for archaic english pronouns like *thee* 'you', *thy* 'your', *thyself* 'yourself' are quite high. Students that are Non-Native speakers of English might come across many difficulties when they encounter English texts which include these kinds of archaic words. How might computer be a help for the student? This paper aims on providing some supports from the perspective of Computer Assisted Language Learning (CALL). It proposes some designs of lexicon transducers by using Local Grammar Graphs (LGG) for auto-convert of the archaic words to PDE in a literature machine readable text. The transducer is applied to a machine readable text that is taken from Sir Walter Scott's Ivanhoe. The archaic words in the corpus can be converted automatically to PDE. The transducer also allows the presentation of the two forms (Arhaic and PDE), the PDE lexicons-only, or the original (Archaic Lexicons) form-only. This will help students in understanding English literature works better. All the linguistic resources here are machine readable, ready to use, maintainable and open for further development. The method might be adopted for lexicon tranducer for another language too.

## 1. INTRODUCTION

Novel is a long written story about people and particular events. The content of a novel describes the characters and events in the novel, while the stylistic of a novel might give impression of the language used during the timeframe of the novel. For instance, the stylistic of John Grisham's *Runaway Jury*, written in 2003, is different from Sir Walter Scott's *Ivanhoe* which was written in 1819. In *Ivanhoe*, readers might find some words that are not in the current use of English, as opposed to *Runaway Jury*; readers can easily identify some archaic words like *thy, thee, nay, couldst* in *Ivanhoe*, while in *Runaway Jury*, a similar occurrence might relatively be difficult to find.

More recently, literary works has emerged not only for the study of literature, but they are often adopted as learning resources for the study of foreign language. One of the most frequently used literary works for the study of English is novel. This is indicated by the existence of *Book Report, Novel Review, Literary Critics* course (or its kind), as mandatory courses for English department students in some universities in Indonesia. This course requires students to read thoroughly the English novels, and to report the comprehension to the lecturer in charge. However, often these students come across many difficulties when they encounter novels which include archaic words; the lexicon choice in these novels is quite distinct from present day English (PDE).

Even though some of the current version of the literary works have already been simplified and written in mostly PDE, there are some archaic words that are still preserved in the novels. On one side, the substitution of these archaic words to PDE might remove some cultural touch in the novels. On the other side, if these archaic words are preserved, it poses some challenges for language learners. These are the main issues that this research is attempting to deal with.

This research aims at proposing a method to cope with the problems from the perspective of Computer Assisted Language Learning (CALL). The research in this paper is machine-readable text driven, and the text is obtained from Sir Walter Scott's *Ivanhoe*, www.online-literature.com[ii], which is also a default text in the corpus processing software employed in this research.

The research can be briefly described as follow. First stage is the identification of archaic words that are used in the novel. The words are examined and classified into a list consisting of two categories based on the patterns of change: regular and irregular. The list then formalized as Machine Readable Dictionary (MRD). MRD is different from printed dictionary. In this paper, we refer to MRD concept, which account for Natural Language Processing (NLP)[iii] in computational. The MRD was enhanced with inflectional MRD, generating a ready-to-use Lexical Resource for the corpus. The corpus itself is processed with UNITEX, one of the local grammar (Gross 1993; Gross 1997) based corpus-processing software, which allows text processing in several languages like French, English, Korean and etc. In processing stage, I designed an auto convert transducer that can perform lexicon recognition and automatic extraction for all archaic words on the corpus text and in turn, gives output in PDE. The locate pattern function in UNITEX enables users to display both input (in this case archaic words) and the output (PDE) at the same time. More details about research procedure will be described in the METHOD section.

This paper is organized in the following way. Section one covers the overview of the research, the aims, the background and the brief summary of the methodology. Section two provides literature review over some essential notions like archaic words and English in current use. It will also briefly describe some essential concepts, such as: CALL, corpus based research, Local Grammar, MRD and Transducer. Section three elaborates the methodology that consists of the description of research corpus and research procedure. Section four provides a useful account on the formalization method to create a Lexical Resource. It also deals with processing and preprocessing of the corpus text. This section will demonstrate how the lexical resource is applied and how to design and to use the transducers for auto converts. Section five concludes the research with a summary and some suggestions for further research.

## 2. LITERATURE REVIEW

### Archaic and Present Day English

Collins Cobuild Dictionary and Oxford Dictionary defines 'archaic' as something that is extremely old or old fashioned. In dictionaries of English, we can find some archaic-labeled entries like *nay, thee, thy,*; it means that these entries are no longer in common use in current English or they come different language era. Consider the chronological development of English language is presented on table 1 from (Baker et al 2006):

**Table 1. Chronological Development of English Language**

| English Language | Year |
|---|---|
| Old English (Anglo Saxon) | up to c.1100 |
| Middle English | c.1100-c.1500 |
| Modern English | c.1500 |
| Present-day English | Now |

Table 1 shows that English language used before 1500 are considered archaic. But still, in the current use of English, we might encounter the use of some archaic words. These words are usually available only on very specific genres like religion or literature; for instance, novels. In order to make these novels comprehensible, the surface representation can be simplified and used for language learning. But we might also need to consider that in these kinds of novels, some of the archaic words are sometimes preserved to provide some historical and cultural touch. Novels that are in machine readable forms are no exception.

**CALL and Corpus Based Research**

The interaction between human and computer has become more and more intensive these days. When managed correctly, this interaction might benefit language learners; for instance, some of the learning resources are available not only in printed form, but also on computer readable format. This is considered time and space saving (when we compare to printed forms resources). Learners can also install some programs for language learning in their computers, and use it independently from the classroom teachers. This is different from the conventional method where students only rely on printed materials and classroom teachers. The process where computers are utilized to support language learning is often referred as Computer Assisted Language Learning (Warschauer 1996). However to what extent computers are used and what kind of material that are used may vary.

Some CALL programs or applications conceive corpus as the reference data. Corpus can simply be defined as the collection of linguistic data[iv] that is machine readable. It allows computer processing for quantitative and qualitative analysis (Baker et al., 2006). Some of the corpus can be accessed on line like British National Corpus (BNC), English WordNet, Korean Lexicon (KorLex), Malay Concordance Project and etc. There are also some software which are dedicated for corpus management and corpus processing such as INTEX[v], UNITEX[vi], NOOJ[vii] etc.

In this research, I use UNITEX, a Local Grammar (Gross 1993; Gross 1997) based corpus processing software. Local Grammar based software have successfully been used in some linguistic and educational research, such as: the recognition of Korean proper names (Nam and Choi 1997), Multiword Annotation in French corpus (Laporte et al 2008), Named entity extraction in Arabic (Traboulsi 2009), automatic generation of Korean language exercise items for Indonesians (Prihantoro 2011) and etc.

**MRD, Local Grammar and Transducer**

Lexical resource is a term used by UNITEX which refer to Machine Readable Dictionary (MRD). MRD is distinguished from published dictionary (printed form). Entries and all information in MRD are machine (Computer readable). However, we must distinguish the MRD for human and for computer. MRDs for human are printed form MRDs which are made electronic like CD ROM or portable dictionary. MRD in this research is meant for computer processing. Here we can understand that the two MRD types are designed for different purpose and processed in different way. Computer MRD can be used to perform NLP tasks such as: automatic extraction, information retrieval, recognition, lexicon generation and etc. MRDs are usually designed for specific language, such as English WordNet, Korean KorLex and DECO, Indonesian WordNet (In progress).

The notion of Local Grammar is proposed by Gross (1993 & 1997). The formalization of Local Grammar can describe and process local linguistic phenomena, which is computer readable, by using Finite State Automaton (or Finite State Transducers). Zilberstein (1993) has created a user friendly formalism tool for local grammar which is called Local Grammar Graph (LGG). LGG can perform some natural language processing tasks over a corpus text or by making reference to the applied lexical resource, another LGG (sub graph) or the combination of two.

Automaton (Finite State) refers to an LGG that has no output; Transducer (Finite State) refers to a mechanism where an LGG can be used to generate output. It is sometimes referred as LGG transducer as it generates a finite set of output. For instance, in figure 1, the expression recognized by the graph is a determiner *the* plus any noun, which constitute an NP. If required, this information can be attached in the concordance display (result). Consider figure 1 and the concordance in result in figure 2 from Prihantoro (2011b):

**Figure 1. Transducer Sample**

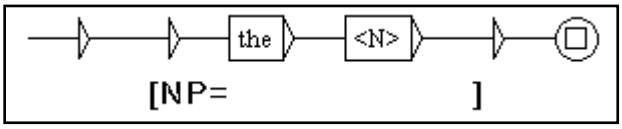



**Figure 2. Concordance Sample**

own situation, and [NP=the appearance] which he ma
e slumbered, under [NP=the appearance] of sullen c
take a turn round [NP=the back] o' the hill to ga
y, by the event of [NP=the battle] of Hastings, an
have mentioned in [NP=the beginning] of the chapt
rt of the cap that [NP=the bells] were attached; w
yet more close to [NP=the body], it was gathered

The LGG in figure 1 is a transducer type. It assigns [NP =] to every string composed of [the + N]. The result of automatic extraction of NPs in an English Corpus is the concordance shown in figure 2. All the tokens after 'the' is tagged with <N> reference code in the dictionary. If the LGG is not transducer type, it is limited only in the extraction without assigning [NP=] output.

The interaction of Corpus, MRD and LGG in this research can be used to support CALL in account for better understanding of archaic and present day English. The information concerning to what extent these linguistic resources interact, and how the research procedure is carried out, are presented in the very next sections of this paper.

## 3. METHOD

### Research Corpus

The machine readable text for this research is obtained UNITEX default English corpus, Sir Walter Scott's Ivanhoe. from on line repository of English Literary works. It is a novel by Sir Walter Scott, which takes setting in the 12$^{th}$ century England. The novel itself was written in 1819. The machine readable version of the novel can also be obtained from www.online-literature.com. It has totally 44 chapters.. After preprocessed by UNITEX, the corpus statistic shows 2606 sentence delimiters, 186877 tokens, and 9299 types (different tokens), 13333 simple lexical entries and 274 compound lexical entries.

### Research Procedure

The research began with the manual identification of archaic words used in the novel. The equivalences in PDE are written with reference to published Collins Cobuild Dictionary (2001). The result is a list which consists of archaic-PDE words. The list was then formalized with entry line formalism to generate a new Machine Readable Dictionary (MRD) that is used as lexical resource in UNITEX. The corpus (*Ivanhoe*) is converted to .txt for this is readable word processing format in UNITEX. The MRD is applied to the corpus text in preprocessing stage. After that, some Finite State Transducers LGGs are designed with reference to the lexical resource. These transducers are employed for pattern matching operation. The concordance will list the target lexicon in the center. Users might want to choose to display both forms (Middle and PDE), or converting the Archaic lexicons to PDE.

## 4. SYSTEM APPLICATION AND PRESENTATION OF THE RESULT

### Findings: Identification of Archaic English

The findings indicate that the archaic words in the corpus can be divided into two according to the patterns of change: regular and irregular. There are some patterns for regular changes. For instances, the equivalence of *couldst* (archaic) is *could* in PDE, which can historically be illustrated as *couldst>could*. In this way, it has lost the *–st* suffix. Another instance, *lackest* has lost its *-est* in PDE. Some other words undergo more complex processes. Consider *carriest>carry*, where the process might be described as the deletion of *-iest* and the addition of *-y*. Another instance is *curst>curse,* where it has undergone the deletion of *-t* and the addition of *-e*.

On the contrary to regular change, words that are classified into irregular change have no predictable pattern, for instances: *nay>no, ay>yes, hath>has, hast>have* and etc. Another instance is *You* in PDE, which has several equivalences in the archaic forms like *ye* [+PLURAL], *thou* [+SINGULAR], *thee* [+SINGULAR|OBJECT VERB|OBJECT of PREP]. Consider some of the examples in table 2:

**Table 2. Some Comparative Samples of Middle to PDE**

| Regular | | Irregular | |
|---|---|---|---|
| **Archaic** | **PDE** | **Archaic** | **PDE** |
| *Carriest* | *Carry* | *nay* | *No* |
| *Lackest* | *Lack* | *ay* | *Yes* |
| *Couldst* | *Could* | *thee* | *You* |

## MACHINE READABLE DICTIONARY (Lexical Resource)

The archaic entries must be distinguished from other entries as archaic entries. In this paper, the code *Q* is assigned to mark archaic entries. Note that this is not absolute since the one who designs the MRD might use another code as long as they do not conflict with other codes (in the same dictionary). The code is assigned by an LLG called 'inflection[viii]' graph (morphology). The resulting entry lines are combined with previously existing MRD (PDE). The file compressed into .bin for better performance. This file is the lexical resource that we are going to use for preprocessing. We will start by discussing non-inflected form MRD.

The format of MRD consists of inflected and non inflected (base) forms, where in this research, the inflected form is set for archaic words, and the base form is set for PDE. This is important since the system will recognize the inflected form (archaic) as the surface form, and use the base form to generate PDE lexicon. For the irregular patterns, the entry lines must be written manually. However, for the regular ones, the inflected forms can be obtained by automatic inflection of the base form. For instance, *lackest* is considered the inflected form of *lack*. Consider the format of the base form MRD in figure 2:

**Figure 2. Some Entries for Base Form MRD**

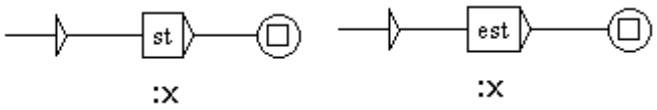
```
argue,Q1+V
be,Q2+V+Aux
bear,Q2+V
can,Q1+Aux
carry,Q3
could,Q1+V+Aux
curse,Q4+V
dare,Q1+V
deserve,Q1
did,Q1+V+Aux
do,Q1+V+Q
do,Q1+V+Aux
know,Q2+V
lack,Q2+V
make,Q1+V
may,Q2+V+Aux
may,Q1+V+Aux
might,Q1+V+Aux
please,Q1+V
refuse,Q1+V
said,Q1+V+Past
say,Q1+V
say,Q2+V
see,Q1+V
seem,Q2+V
shall,Q4+V
should,Q1+V+Aux
speak,Q2+V
swear,Q2+V
think,Q2+V
```

The entry line <can,Q1+V+Aux> is composed of : entry *can*, archaic lexicon code <Q>, and inflection code <2>. Other additional codes may exist like grammar codes N (Noun), V (Verb), Conj (Conjunction) or Aux (Auxiliary).

There are four lexicon codes assigned to the Middle-PDE words that change regularly <Q1>,<Q2>,<Q3> and <Q4>; The first two codes are assigned for direct concatenation types, and <Q3> and <Q4> are assigned for non direct concatenation types.

In this base form MRD, the entries are all in PDE. These entries are going to be inflected by inflection graphs (morphology), which comply with inflection codes in the dictionary format. Note that entries with different inflection codes are inflected in different ways. Consider figure 3:

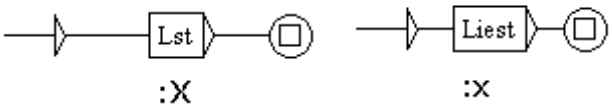
**Figure 3. Q1 and Q2 Inflection Graphs: Direct Concatenation**




The two graphs in figure 2 concatenate *-st*, and *-est* directly to PDE entries generating archaic entries as inflected forms. For instance *can* will be inflected by Q1 graph resulting on *canst* as the inflected form. Another example is *seem* that will be inflected by Q2 graphs and in turn, *seemest* will be generated. Note here that the change is reversed from the historical process described in the finding section. This is because, we want to recognize the Archaic lexicon as the surface inflected form and the PDE as the base form.

Consider also some other inflecting operations where we cannot perform direct concatenation; for instance, type 3 lexicons like *curse-curst* or type 4 lexicon like *carry-carriest*; deletion is required for each. These types of entries are inflected by Q3 and Q4 graphs as presented in figure 4:

**Figure 3. Q3 and Q4 Inflection Graphs: Deletion**




Inflection graph Q3 removes one letter from the left and insert <-st>. Therefore, for entry *curse*, it deletes *e* and add *–st* generating *curst* as the inflected form. Inflection graph Q4 deletes one letter from the left and inserts *–iest* generating inflected form *carriest* from a base form entry *carry*. Consider figure 4:

**Figure 4. Inflected Form MRD for Regular Pattern**

```
arguest,argue.Q+V:x
beest,be.Q+V+Aux:x
bearest,bear.Q+V:x
canst,can.Q+Aux:x
carriest,carry.Q:x
couldst,could.Q+V+Aux:x
curst,curse.Q+V:x
darest,dare.Q+V:x
deservest,deserve.Q:x
didst,did.Q+V+Aux:x
dost,do.Q+V+Q:x
dost,do.Q+V+Aux:x
knowest,know.Q+V:x
lackest,lack.Q+V:x
lackest,lack.Q+V:x
makest,make.Q+V:x
mayest,may.Q+V+Aux:x
mayst,may.Q+V+Aux:x
mightst,might.Q+V+Aux:x
pleasest,please.Q+V:x
refusest,refuse.Q+V:x
saidst,said.Q+V+Past:x
sayst,say.Q+V:x
sayest,say.Q+V:x
seest,see.Q+V:x
seemest,seem.Q+V:x
shalt,shall.Q+V:x
shouldst,should.Q+V+Aux:x
speakest,speak.Q+V:x
swearest,swear.Q+V:x
thinkest,think.Q+V:x
wert,were.Q+V+Aux:x
whilest,while.Q+Conj:x
wouldst,would.Q+V+Aux:x
```

In the inflected form MRD presented by figure 4, each entry line is composed of <archaic lexicon>, <PDE lexicon>, <Q: Archaic Code>, <V,N, Aux: POS codes>, <x: regular change code>.

The format for lexicons with irregular changes is almost similar. It is distinguished by the inflection code *y*, instead of *x*, to indicate irregular changing pattern. Consider the inflected form MRD for irregular pattern in figure 5:

**Figure 5. Inflected Form MRD for Irregular Change**

```
hast,have.Q+V+Aux:y
hath,has.Q+V+Aux:y
tempest,storm.Q+N:x
repast,meal.Q+N:y
ye,you.Q+N+Pro:y
thy,your.Q+N+Pro:y
thou,you.Q+N+Pro:y
thee,you.Q+N+Pro:y
thyself,yourself.Q+N+Pro:y
thine,your.Q+N+Pro:x
nay,no.Q+N:y
ay,yes.Q+N:y
thither,there.Q+N:y
sirrah,sir.Q+N:y
doth,do.Q+V+Aux:y
spake,speak.Q+V+Q:y
ere,before.Q+P:y
```

Figure 5 presents the inflected form MRD for irregular change from middle to PDE. Unlike the regular one, this MRD cannot be generated automatically. Instead, it must be written manually. One that distinguishes the format from the regular MRD is code *y* where it indicates irregularities of change from middle to PDE. The code is distinguished in order to ease users when they are required to perform separate extraction (regular <:x>/irregular<:y>).

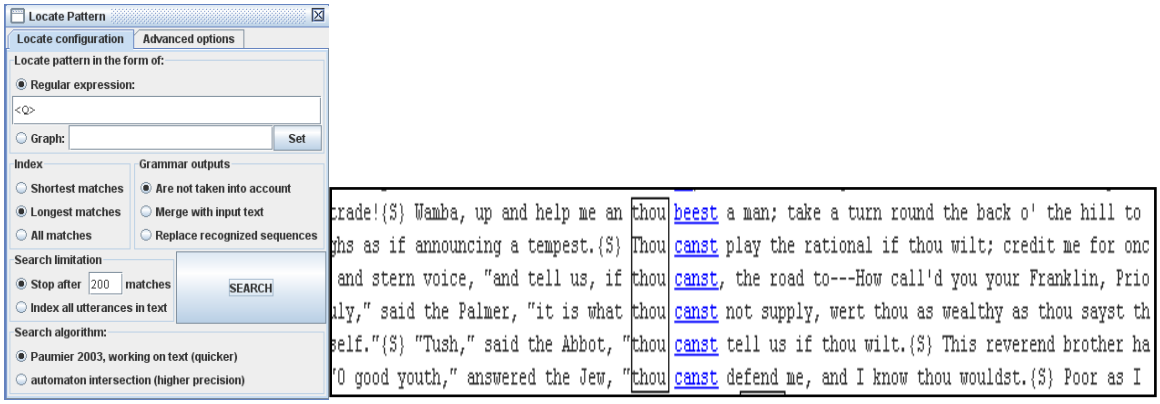
The entries from inflected form MRD are copied to standard MRD resulting on the enhanced version. Therefore, the entries in the enhanced version of base form MRD (Inflected form) now consist of modern and archaic entries. This enhanced MRD is applied in the text corpus on preprocessing stage. It results on default annotation on PDE lexicons and additional annotation for Archaic lexicons. Consider wordlist in figure 6:

**Figure 6. Wordlist for Enhanced MRD**

```
thee,you.Q+N+Pro:y
theelin,.N:s
theelins,theelin.N:p
theelol,.N:s
theelols,theelol.N:p
theft charge,.N+XN+z1:s
theft charges,theft charge.N+XN+z1:p
theft loss,.N+XN+z1:s
theft losses,theft loss.N+XN+z1:p
theft victim,.N+XN+z1:s
theft victims,theft victim.N+XN+z1:p
theft,.N:s
theftless,.A
```

After the preprocessing the tokens are already annotated and we might perform the extraction of the Archaic words in the text with a single query in the regular expressions by using <Q> as the code that has been set in the enhanced MRD. The extraction is performed by regular expression as it is presented in figure 7:

**Figure 7. Retrieval with Regular Expression**



From the concordance display we can observe the contexts where the words are used. However, the consequence for using regular expression is that we might not be able to obtain output (lexicon generation); this does not comply with our purposes.

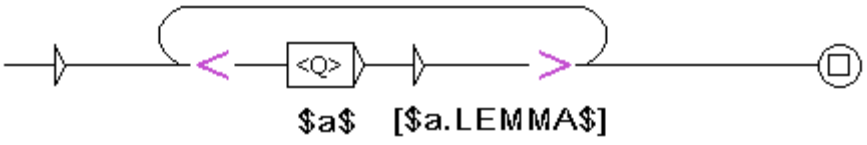
Another issue that must be resolved is that the extraction limits the scope for only one token. In fact there are some strings where the archaic words are composed of several concatenated tokens. For instance, in figure 7, the concordance displays the extraction of *beest* and *canst* separately from *thou*. In fact, strings composed of more than one token should be extracted together. The solution for these two problems (Lexicon Generation and Extraction) is LGG transducer type.

## LOCAL GRAMMAR GRAPHS AS TRANSDUCER

Transducer is an LGG mechanism for pattern matching operation that allows both input (recognition) and output (Lexicon Generation). Here I built transducers that allow recognition of Archaic words and generate output as PDE. The transducers will: make reference to the lexical resource (MRD), extract the codes (not grammar codes). and display the equivalence with the archaic inflection code <:*x*> for the regular change, and <:*y*> for the irregular change.

It is also possible to manipulate concordance to display the two forms (both Middle and PDE) or to change directly to PDE words (Archaic words are removed completely) . The first option seems to be better because the students might be able to compare the Archaic forms and the PDE forms. They can also compare the sentence context where the words are used. From the inflection codes (<:*x*> or <:*y*>), students might also know that some words have relatively predictable patterns, while some others are not. Consider figure 8 which is the graphical representation of the transducer LGG used in this research:

**Figure 8. LGG Transducer for Lemma Output**




In the transducer presented in figure 8, there is a loop which is useful to detect recursion of the tokens annotated with code <Q> (archaic words). Therefore, strings composed of archaic tokens (more than one token) might be extracted. Consider figure 9 and 10:

**Figure 9. Locate Pattern Operation (Input Only)**

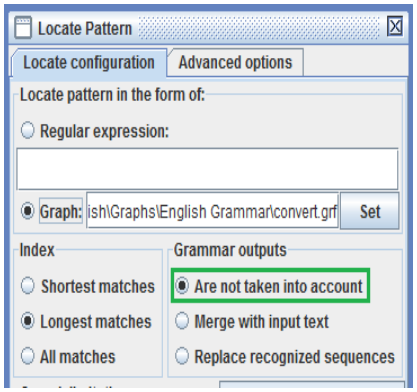



**Figure 10. Concordance (Input Only)**

Enough," said the Disinherited Knight, "thou knowest my promise."{S} "Nay, for that matter," sa
t confession upon Good Friday eve."{S} "Thou knowest best thine own privileges," said DeBracy.{
the Templar, "what hast thou to fear?---Thou knowest the vows of our order."{S} "Right well," s
I possessed---Yet I can tell thee what thou lackest, and, it may be, supply it too.{S} Thy wis
to discover who or what we are; for, if thou makest such an attempt, thou wilt come by worse fo
e."{S} "By my faith," said the knight, "thou makest me more curious than ever!{S} Thou art the
t far from the town of Sheffield, where thou mayest easily find many of thy tribe with whom to
ls bad not dizzied thine understanding, thou mightst know Clericus clericum non decimat; that i
ercy! good fellow," cried Prince John, "thou pleasest me---Here, Isaac, lend me a handful of by
l me to bend or to draw my bow."{S} "If thou refusest my fair proffer," said the Prince, "the P
or abstain from taking by violence what thou refusest to entreaty or necessity."{S} "Stand back
"I will remain beside my prize.{S} What thou sayst is passing true, but I like not the privileg
lgrim, "thou art but a Saxon fool."{S} "Thou sayst well." said the Jester; "had I been born a N
nst not supply, wert thou as wealthy as thou sayst thou art poor.'{S} "As I say?" echoed the Je
n which, if thou best really that which thou seemest, thou mayst take an honourable part.{S} A
lan, wilt thou not?--Thou hast nothing, thou seest, to fear from my interference."{S} "No," rep
bert, at what ransom were they held? ---Thou seest thou canst not deceive me."{S} "My master,"
leaders of the banditti.{S} "It is time thou shouldst leave us, Sir Maurice," said the Templar
r wealth, to take thy seat, honoured as thou shouldst be, and shalt be, amid all in England tha
r, thou wilt return them safely--unless thou shouldst have wherewith to pay their value to the
so readily?"{S} "I think," said Gurth, "thou shouldst be best able to reply to that question."{
"But for my purpose," said the yeoman, "thou shouldst be as well a good Englishman as a good kn
{S} As for thy threats, know, holy man, thou speakest to one whose trade it is to find out dang
e tithe of that huge sum of silver that thou speakest of."{S} "I am reasonable," answered Front
somewhat heated, "thou knowest not what thou speakest---His neck and limbs are his own, but his
."{S} "By St Dunstan," answered Gurth, "thou speakest but sad truths; little is left to us but
"Bethink thee, man," said the Captain, "thou speakest of a Jew---of an Israelite,---as unapt to
."{S} "Gurth," said the Jester, "I know thou thinkest me a fool, or thou wouldst not be so rash
-refuse to employ it, Wilfred dies, and thou thyself art not the nearer to freedom."{S} "Thy la

Figure 9 presents the pattern matching box, where LGG based pattern matching is preferred over regular expression. We may observe on concordance in figure 10, that the loop extracts both individual tokens and it also allows the extraction of strings composed of multiple Archaic tokens, for instances: *thou knowest* 'you know' (two tokens), *thou thyself art* 'you yourself are' (three tokens), *thou sayst thou art* 'you say you are' (four tokens).

The transducer in figure 8 is manipulated so that it allows recognition of Archaic words in the text, and at the same time, it can generate output in PDE with the following formula [$.a.LEMMA$]. With this formula, the LGG will make reference to the lexical resource to generate output from the base form. Please consider locate pattern box in figure 11, and the concordance display in figure 12:

**Figure 11. Locate Pattern (Input+Output)**

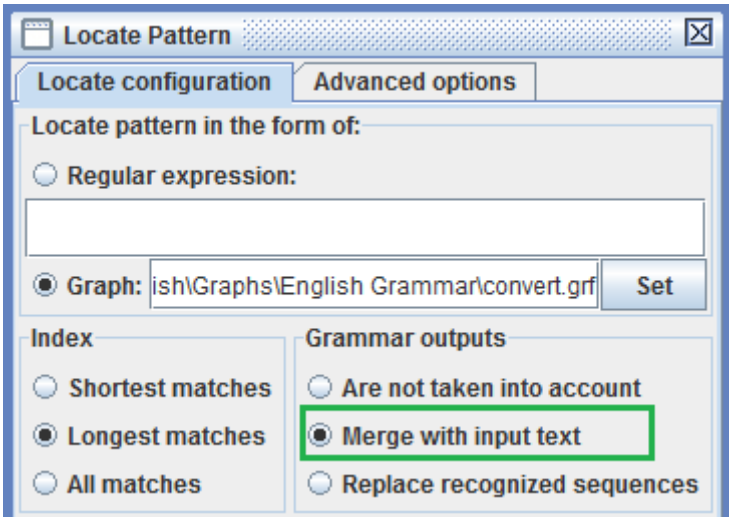



**Figure 12. Concordance (With Output)**

"By my faith," said the knight, "thou[you] makest[make] me more curious than ever!{S} Th
rom the town of Sheffield, where thou[you] mayest[may] easily find many of thy tribe wit
not dizzied thine understanding, thou[you] mightst[might] know Clericus clericum non dec
ood fellow," cried Prince John, "thou[you] pleasest[please] me---Here, Isaac, lend me a
bend or to draw my bow."{S} "If thou[you] refusest[refuse] my fair proffer," said the P
ain from taking by violence what thou[you] refusest[refuse] to entreaty or necessity."{S
remain beside my prize.{S} What thou[you] sayst[say] is passing true, but I like not th
"thou art but a Saxon fool."{S} "Thou[you] sayst[say] well." said the Jester; "had I bee
supply, wert thou as wealthy as thou[you] sayst[say] thou[you] art[are] poor.'{S} "As I
, if thou best really that which thou[you] seemest[seem], thou mayst take an honourable
lt thou not?--Thou hast nothing, thou[you] seest[see], to fear from my interference."{S}
t what ransom were they held? ---Thou[you] seest[see] thou[you] canst[can] not deceive m
of the banditti.{S} "It is time thou[you] shouldst[should] leave us, Sir Maurice," said
h, to take thy seat, honoured as thou[you] shouldst[should] be, and shalt be, amid all i
wilt return them safely--unless thou[you] shouldst[should] have wherewith to pay their
ily?"{S} "I think," said Gurth, "thou[you] shouldst[should] be best able to reply to tha
r my purpose," said the yeoman, "thou[you] shouldst[should] be as well a good Englishman
for thy threats, know, holy man, thou[you] speakest[speak] to one whose trade it is to f
of that huge sum of silver that thou[you] speakest[speak] of."{S} "I am reasonable," an
t heated, "thou knowest not what thou[you] speakest[speak]---His neck and limbs are his
By St Dunstan," answered Gurth, "thou[you] speakest[speak] but sad truths; little is lef
k thee, man," said the Captain, "thou[you] speakest[speak] of a Jew---of an Israelite,--
Gurth," said the Jester, "I know thou[you] thinkest[think] me a fool, or thou wouldst no
to employ it, Wilfred dies, and thou[you] thyself[yourself] art[are] not the nearer to

Figure 12 presents pattern matching box, where <merge with input text> option is selected to perform lexicon generation. The result is displayed by concordance as it is displayed in figure 12. There, each of

the archaic words are adjunct to the equivalent in PDE marked by square brackets as in <*thou* [you] *sayst* [say] *thou* [you] *art* [are]>. By using this option, students are able to know both the Middle and PDE words.

Complete removal of the archaic words is also possible. If this option is selected, then none of the archaic word is shown on the concordance. For this purpose, when performing pattern matching, <replace recognized sequence> must be opted. The consequence is, it will completely remove all the archaic words, and display only the PDE equivalences. Please consider figure 13 and figure 14:

**Figure 13. Pattern Matching Box (Output only)**

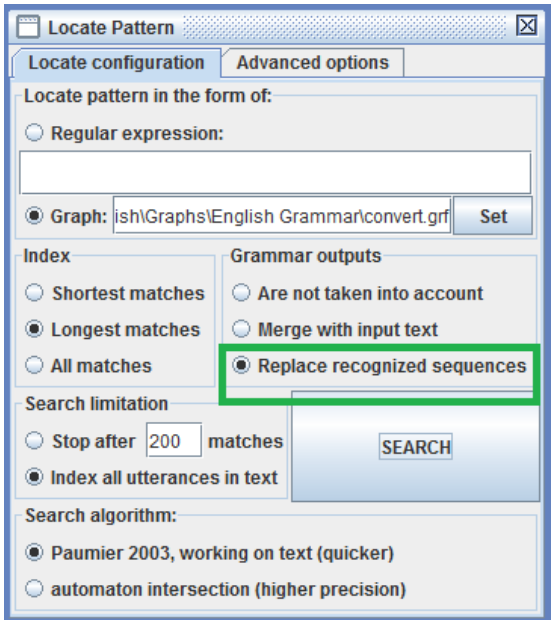



**Figure 14. Concordance Display (Output only)**

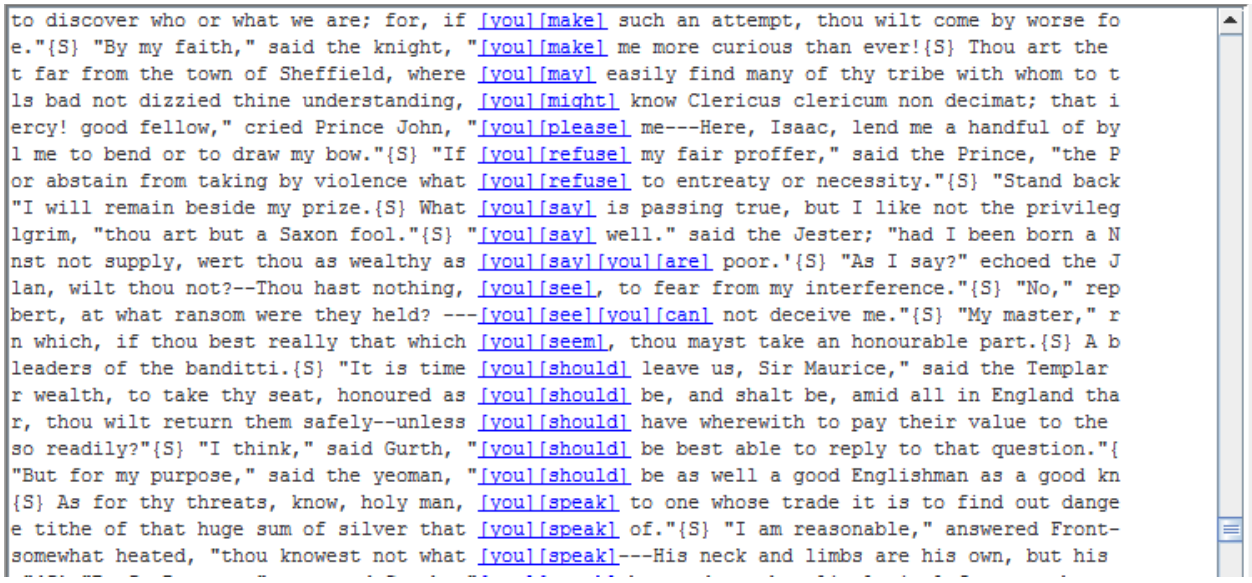


Figure 13 demonstrates how to remove the archaic words completely, and replace it with PDE. The process can be described as follow: The LGG transducer performs pattern matching operation by detecting the archaic words on the text. Then it generates PDE words as output, and at the same time, it removes the archaic words from the concordance display. We can observe the result in figure 14.

## ARCHAIC WORDS IN DIRECT SPEECH

One of the interesting findings by the LGG transducer is that it managed to detect that archaic words are used in quotation. In other words, they are all composed in direct speech sentence constructions. This means that these archaic words come from the characters in the novel, and not from the narrators. Consider figure 15:

**Figure 15. Archaic Words and Direct Speech from the Characters**


The LGG transducer is manipulated to recognize quotation marks, marked by QUOTE=>. We can observe that strings composed of archaic words, marked by MID-->, are located in the direct speech.

Some of the archaic words have no longer been used at all; some others are used in very specific works like poems or drama scripts or novels. The language simplification of English novels aims on enlarging the scope of readability since the language used at the time of writing might be different from PDE. It also serves the purpose of education since the novels are used as one of the learning materials for English learners. At the same time, complete removal of these archaic words might also remove the cultural touch at the time when the novel is written. Therefore, to accommodate these issues, the narration is changed to PDE but some archaic words are preserved in the utterance from the characters (direct speech). But the remaining question is why they always come from the characters, and not the narrator. This is a non-computational issue, but it is quite interesting to discuss from other perspectives.

## 5. CONCLUSION

This paper has demonstrated how Machine Readable Dictionary (MRD) and LGG Transducers might be employed in CALL to help students in understanding literary texts that contain archaic words. These computer readable resources work on the research corpus to perform recognition, automatic extraction and lexicon generation which might be summarized in the following details:

1. Recognition of archaic words, both individual tokens and multiple strings.
2. Automatic extraction of archaic words and the assignment of the equivalences in PDE
3. Substitution of archaic words with their equivalences in PDE

The machine readable resources of this research are maintainable and open for further development. For further studies, the archaic entries of MRD might be improved. It suggests in depth studies of archaic entries are used in the literary works. The size of corpus can also be improved so that more entries can be covered. Some of the additional grammar and semantic tags can also be assigned to enrich linguistic information regarding the entries. When the MRD is improved, the LGGs can also be manipulated to tighten or to loose the constraints.

## REFERENCES

Baker, P., Hardie, A., and Mecenery,T.(2006). A Glossary of Corpus Linguistics. Edinburgh University Press: Edinburgh

Fellbaum, C. (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Gross, M. (1993). Local Grammars and their Representation by Finite Automata. Data, Description, Discourse. Papers on the English Language in honour of John McH Sinclair, M. Hoey (Ed.) (1993) 26-38"

_____. (1997). The Construction of Local Grammars. Finite-Sate Language Processing. The MIT Press.

Horobin, S., & Smith, J.(2002). Introduction to Archaic. Edinburgh University Press: Edinburgh

Laporte, A., Nakamura, T. and Voyatzi.S. (2008). A French Corpus Annotated for Multiword Nouns. Towards a Shared Task for Multiword Expressions (MWE 2008). P27-31

Nam J.-S. & K.-S. Choi. (1997). "A Local Grammar-based Approach to Recognizing of Proper Names in Korean Texts," in Proc. Workshop on Very Large Corpora, ACL, Tsing-hua Univ. and Hong-Kong University of Science and Technology, pp. 273-288.

Paumier, Sebastien. (2008). Unitex Manual. Université Paris-Est Marne-la-Vallée: Paris

Prihantoro. (2011a) A Comparative Study of Korean and Indonesian Noun Phrases with Numerals: Building A Machine Readable Linguistic Resource. Master Thesis. Hankuk University of Foreign Studies.

_____.(2011b) Local Grammar based Auto-Prefixing Model for Automatic Extraction in Indonesian Corpus (Focus on Prefix meN-). Proceeding of KIMLI, International Congress of MLI, October 2011: UPI Bandung

Silberztein, M. (1993). Dictionnaires électroniques et analyse automatique de textes : le système INTEX. Masson Ed.: Paris

_____. (1997). The Lexical Analysis of Natural Languages. Finite-Sate Language Processing. The MIT Press.

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

## CONSULTED DICTIONARIES AND CORPUS

*Korlex 1.5*: NLP Lab, Busan University. Retrieved from http://corpus.fr.pusan.ac.kr/korlex/

*Collins Cobuild Dictionary* 3.1 (Off Line Version)

*British National Corpus*, Retrieved from http://www.natcorp.ox.ac.uk/

*Malay Concordance Project*, Australian National University. Retrieved from http://mcp.anu.edu.au/

*Oxford English Dictionary*, Retrieved from http://oxforddictionaries.com/

### Endnotes

[i] Some abbreviation used in this research: CALL (Computer Assisted Language Learning), MRD (Machine Readable Dictionary), LGG (Local Grammar Graph), PDE (Present Day English)

[ii] This website is a repository of literary works, indexed by name of popular authors like William Shakespeare, Oscar Wilde, Charlotte Bronte etc.

[iii] NLP is a term that is widely used in computer science for computational linguistics.

[iv] It is widely known that corpus is the collection of texts. However, these days there are also some corpora for spoken language or often referred as speech corpora, like Santa Barbara Spoken Corpus (http://www.linguistics.ucsb.edu/research/sbcorpus.html) , Buckeye Speech Corpus (http://buckeyecorpus.osu.edu/) and spoken language corpora on Multilingualism Center (http://www.exmaralda.org/corpora/en_redirect.html) . That is why, I do not use 'texts' but linguistic data as it can cover both spoken and written forms.

[v] http://www.nyu.edu/pages/linguistics/intex/

[vi] http://igm.univ-mlv.fr/~unitex/

[vii] http://www.nooj4nlp.net/pages/nooj.html

[viii] This is a technical term in UNITEX. Some may disagree as word formation of inflection might be distinguished from derivation. In UNITEX however, inflection LGG refers to the attachment of any affix/es regardless of they are morphologically considered as inflection or derivation.