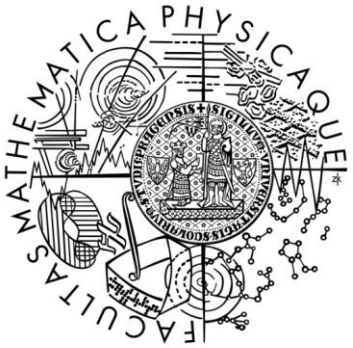
Charles University in Prague
Faculty of Mathematics and Physics

Master's Thesis



Septina Dian Larasati

Machine Translation on Related Austronesian Languages

Institute of Formal and Applied Linguistics

Supervisor: RNDr. Vladislav Kuboň, Ph.D.

Study Program: Computer Science
Mathematical Linguistics

2010

[This page is intentionally left blank]

I would like to thanks to my supervisor Dr. Vladislav Kuboň, for his guidance and support throughout the work. I am also thankful for the opportunity and guidance from the Apertium Community especially for Francis Tyers that helped me on technical detail problems. And I also would like to thanks Dr. Mans Hulden for his help on some Foma implementation details. Thank to the native speakers who gave me feedbacks on the evaluation. And last but not the least thanked, a lot of people around me that support me during my Master studies that I can't mentioned individually.

I certify hereby that this diploma thesis is all my work and that I used only the cited literature. The thesis is freely available for all who can use it.

Prague, August 2010

Septina Dian Larasati

[This page is intentionally left blank]

Table of contents

[This page is intentionally left blank]

Title: Machine Translation on Related Austronesian Languages
Author: Septina Dian Larasati
Department: Institute of Formal and Applied Linguistics
Supervisor: RNDr. Vladislav Kuboň, Ph.D.
Supervisor's e-mail address: vk@ufal.mff.cuni.cz

Abstract: This thesis presents the development of an MT system between Indonesian and Malaysian. The system uses a method of almost a direct translation exploiting the similarity of both languages. This method was previously used on a number of language pairs of European languages. The thesis also elaborates the attempts to make language resources from scratch since the languages are under-resourced.

Keywords: machine translation; related languages; direct translation; morphology; hybrid method; finite-state transducers; under-resources language

Chapter 1

Introduction

Machine Translation (MT) is one of Natural Language Processing (NLP) disciplines that has been extensively researched and made into practice. Several MT approaches have been implemented and tested between different language pairs. This thesis focuses on one of the approaches that is *Rule-Based Machine Translation* (RBMT) and more particularly RBMT between closely related language pair. There has been a long history of experiments within this area of interest on different language group such as in Slavic, Scandinavia, Turkic, Celtic, and Romance. So this thesis adds up the list by adding an experiment in another language group that is Malay. The experiment is done for Indonesian (`id`) and Malaysian (`ms`) which both are Austronesian languages. The languages under question are closely related grammatically and syntactically with a rich and different kind of morphological mechanism as compare to European languages.

1.1 Indonesian-to-Malaysian MT system

The close relatedness in terms of lexicography, grammatically, or any other linguistics aspects between the languages can be seen as an ideal setting for a *shallow-transfer method* in RBMT. This shallow-transfer approach had shown a viable result on some close related language pairs in various language groups, one among many reasons is because the languages mostly shares the same features or paradigms where most of the differences can be captured accurately by transformation rules. Indonesian and Malaysian are similarly rich in morphology and highly inflective and having a quite similar grammar with several lexicography differences.

If we compare another approach such as *Statistical Machine Translation* which exploits a big number of language resources, in general the approach will not perform as best as the approach can offer since it is applied to under-represented languages which has language resources availability problem. While in the other hand a relatively small size but accurate language resources somehow enough for the shallow-transfer method to come up with a good result.

Looking at the close-relatedness and resource availability of Indonesian (`id`) and Malaysian (`ms`), the shallow-transfer method currently seems to be a perfect choice. And looking back to several previous experiments, this method has been applied in different language groups, such as

- Slavic languages, e.g. between Czech and Slovak
- Scandinavian languages, e.g. between Norwegian and Swedish
- Turkic languages, e.g. between Turkish and Crimean Tatar
- Celtic languages, e.g. between Irish and Scottish Gaelic
- Romance languages, e.g. between Portuguese and Spanish

This `id-ms` MT system development concentrates on different areas. Those areas are the treatment of the morphological phenomena, the creation of monolingual dictionaries and bilingual dictionary, and the transfer rules. Mainly the work falls in the morphology area where we try to make morphological tools, analyzer and generator, from scratch to reach a certain level of morphological phenomena coverage. The other big challenge is the development of the dictionaries, where it also made from scratch with some heuristics and later also includes a hands-on job on the tailoring. We did several experiments to make the dictionaries in a reasonable size to support the system.

The development is using an MT platform that is Apertium, a free/open source platform for developing rule-based machine translation system. The division of the work is well accommodated by that platform, although throughout the implementation it encountered several challenges, since the morphology technology that is used is newly integrated in Apertium.

1.2 Structure of the Thesis

The thesis can be seen roughly split into three major parts. Chapter 2 will be the introduction to the language under question in linguistics perspective, to give some idea to non-native audiences. Chapter 3, 4, 5, and 6 will go more in detail about each sub-problem of the work. Chapter 7 and 8 concludes the work with Evaluation and Conclusion discussion. This thesis also includes the complete morphological analyzer/generator design described in *finite state diagrams* available in Appendix A.

The linguistics knowledge of the language that is needed for the implementation of the MT system is elaborated in Chapter 4. It also includes examples of the inflection that occurs in the language.

Chapter 3 is the introduction of the MT platform and a brief introduction to the language resources. Chapter 4, 5, and 6 go further into implementation detail. Chapter 4 is dedicated to describe the implementation of the *finite-state transducers* (FST) , starting from the tool selection process, followed by the basic notation used in the design, and the design itself including the FST structures and tagset that is used. The creation of the dictionaries is described in Chapter 5, starting from creating a comparable corpus as a starting point and then explains several attempts on making the dictionaries. Chapter 6 describes the transfer between the languages and the additional processing needed because of the limitation of new tool integration to the platform.

The evaluation and the further discussion of the method applied are available in Chapter 7 and 8.

Chapter 2

Language Pair Typology

The Austronesian family language is one of the major family languages. It spreads throughout islands in Southeast Asia and Pacific region. The related language pair in question in this work, Indonesian (id) and Malaysian (ms), is composed of two widely spoken Austronesian languages which are variants of Malay language group. Indonesian and Malaysian are both standardized Malay language and are official languages of Indonesia and Malaysia respectively. Indonesian, locally known as *Bahasa Indonesia* (lit. language of Indonesia) was declared as a national language of Indonesia at Youth's Oath in 1928. Indonesian is spoken by approximately 160 million people with only 23 million native speakers. Malaysian, or known as *Bahasa Malaysia* and *Bahasa Melayu* in Malaysia and Brunei, is spoken by 17 million speakers.

These languages are closely related where they mostly share similar features in morphology and syntactic aspects. Nevertheless, both of the languages are slightly mutually unintelligible which may lead to a misinterpretation between both speakers. The typology of the language appears to be very similar which makes non-native speakers not able to recognize the differences. These similarities on morphology and syntactic aspect encourage the application of shallow-transfer method.

The typology of the two languages will only be discussed in syntactic and morphology aspects, while the phonetics, phonology, and also semantic aspects will be disregarded since they are not related to the current work.

2.1 Syntactic Aspect

Both languages have the similar syntactic aspect. Describes here some of those aspects such as word order, words agreement, negation, and tense.

Word Order – Indonesian and Malaysian are not free word order languages. The basic word order for both languages is Subject Verb Object (SVO). Since there is no case marking of the words on both languages, the positions of the words are crucial to identify the participants of the event. Both of the languages are non-pro-drop where the subject of the sentence has to be explicitly stated.

Words Agreement – In both languages there is no word agreement such as subject-verb agreement or adjective-noun gender agreement, except number-noun agreement for quantity defined nouns. The quantity is defined by explicit numerals or a determiner that

represents quantity for example the case of the nouns *buku* (book) and *buku-buku* (books) that follows numerals or a determiner *banyak* (many/much) as shown below

1 *buku, 2 buku*	→	1 book and 2 books
satu buku, dua buku	→	one book and two books
banyak buku	→	many books
**banyak buku-buku*		
**2 buku-buku*		
**dua buku-buku*		

Negation – There are two negations in both languages, namely *tidak* (no) and *bukan* (not). The word *tidak* (no) is used for negating verbs and adjectives while *bukan* (not) is used for rests.

Tense – Both of the language does not have any morphological tense marker. The tenses are given by explicitly stating the time adverbs or an additional modal verb. Given below examples of sentences in present, past, future, continuous, and perfect tenses

(a)	*Aku*		*pergi*	*ke*	*sekolah*	
	(I)		(go)	(to)	(school)	
	I		go	to	school	
(b)	*Aku*		*pergi*	*ke*	*sekolah*	*kemarin*
	(I)		(go)	(to)	(school)	(yesterday)
	I		went	to	school	yesterday
(c)	*Aku*	*akan*	*pergi*	*ke*	*sekolah*	
	(I)	(will)	(go)	(to)	(school)	
	I	will	go	to	school	
(d)	*Aku*	*sedang*	*pergi*	*ke*	*sekolah*	
	(I)		(go)	(to)	(school)	
	I	am going		to	school	
(e)	*Aku*	*sudah*	*pergi*	*ke*	*sekolah*	
	(I)	(already)	(go)	(to)	(school)	
	I	have already gone		to	school	

Figure 2-1: (a) Simple present, (b) simple past, (c) simple future, (d) simple present continues, (d) simple present perfect

2.2 Morphology Aspect

Orthography – Both languages are using the 26 letters of Latin alphabet as in English; therefore it can be easily encoded even in ASCII.

Part-of-Speech (PoS) – There is no special markers for the PoS, although for derived words, the derivation morphemes can be used to identify the PoS of the derived words.

Vocabulary – Both languages share some vocabularies either words with exact meaning but also words with different meaning that can be misinterpreted by both native speakers.

English word	*Indonesian Translation*	*Malaysian Translation*
we	*kita* (including listener)	*kita* (including listener)
we	*kami* (excluding listener)	*kami* (excluding listener)
I you	*awak* (colloquial)	*awak* (colloquial)
software	*piranti lunak*	*perisian*

Table 2.1: Indonesian and Malaysian vocabulary examples

Language Type – Indonesian and Malaysian are considered as synthetic language or agglutinative language to be particular, where the words are composed by more than one morpheme. Although in some cases you can find the basic form such as lemma which can stands by itself as in isolating/analytic languages, such as in Chinese or Vietnamese where most of the words have no inflections. While in other cases the word itself can represent a whole clause. Agglutinative languages in theory ideally have these properties:

1. One morpheme conveys one language feature, for example Indonesian prefix *di-* in the word *dipanggil* (being called) only conveys passive voice, as opposed to English suffix *–s* in the word 'smiles' that conveys 3rd person, singular, and present tense.
2. There is a clear-cut boundary between morphemes.
3. Grammatical processes are expressed through affixes and do not change morphemes forms.

Finnish Example		*Indonesian Example*	
talo	(house)	*panggil*	(call)
talo+***ni***	(my house)	***di***+*panggil*	(being called)
talo+***ssa***	(in the house)	***ku***+***di***+*panggil*	(I am being called)
talo+***ssa***+***ni***	(in my house)	***ku***+***di***+*panggil*+***nya***	(I am being called by him/her)
talo+***ja***	(houses)		
talo+***i***+***ssa***	(in the houses)		

Table 2.2: Word's agglutinating examples in Finnish and Indonesian

Indonesian and Malaysian have similar morphological operations that further on will be discussed in 2.2.1 Morphological Operations, where those operations bring language features that will be discussed in more detail in 2.2.2 Language Features.

2.2.1 Morphological Operations

Here describes the morphological operation that includes morphosyntactic and morphophonemic operation. There are four different morphosyntactic operations that are considered in this work namely affixation, compound word formation, reduplication, and clitics.

Affixation – is an operation of adding morphemes in forms of affixes to the current word form. There are four main categories of affix namely,

1. Prefix, affix which preceded the word, for example *meN-*, *di-*, *peN-*, *per-*, and etc. Most of the prefixes such as *di-* and *per-* simply glued to the current word form without any changes, but prefixes such as *meN-* and *peN-* will have some morphophonemic changes.

Affixation: Prefix with Morphophonemic Rules	*Affixed Word*
meN + ***l****ompat* (to jump) also applies for words starting with m, n, q, r, and w	*me**l**ompat* (jump)
meN + ***a****mbil* (to take) also applies for words starting with e, g, h, i, o, u, and x	*me**nga**mbil* (take)
meN + ***b****akar* (to burn) also applies for words starting with f and v	*me**mb**akar* (burn)
meN + ***p****otong* (to cut) *meN* + ***pr****otes* (to protest)	*me**m**otong* (cut) *me**mpr**otes* (protest)
meN + ***d****orong* (to push) also applies for words starting with c, j and z	*me**nd**orong* (push)
meN + ***t****olong* (to help) *meN* + ***tr****ansfer* (to transfer)	*me**n**olong* (help) *me**ntr**ansfer* (transfer)
meN + ***s****apu* (to sweep) also applies for words starting with s and y	*me**ny**apu* (sweep)
meN + ***k****ejar* (to chase) *meN* + ***kl****arifikasi* (to clarify)	*me**ng**ejar* (chase) *me**ngkl**arifikasi* (clarify)
meN + *tik* (to type) applies on word with only one syllable	*me**nge**tik* (type)

Table 2.3: Prefix Operation Examples with Morphophonemic Rules

2. Suffix, affix which simply followed the word, for example *-kan*, *-an*, and *–i*.

Affixation: Suffix	*Affixed Word*
makan (to eat) + *-an*	*makanan* (food)
minum (to drink) + *-an*	*minuman* (drink)
tempel (to stick) + *-an*	*tempelan* (sticker)

Table 2.4: Suffix Operation Examples

3. Infix, affix which placed inside the word, for example *-el-*, *-em-*, and *-er-*. Currently infix is rarely used.

Affixation: Infix	*Affixed Word*
-el- + *tunjuk* (to point)	*telunjuk* (point finger)
-em- + *jari* (finger)	*jemari* (fingers)
-er- + *gigi* (teeth, gear)	*gerigi* (tooth, gear)

Table 2.5: Infix Operation Examples

4. Circumfix, affixes which wrap the word with a combination of prefix and suffix which together treated as whole, for example *per-an*, *ke-an.*

Affixation: Circumfix	*Affixed Word*
per- + *coba* (to try) + *-an*	*percobaan* (experiment, trial)
ke- + *gembira* (happy) + *-an*	*kegembiraan* (happiness, excitement)

Table 2.6: Circumfix Operation Examples

Compound Word Formation – new words can also formed by compounding two or more different words together. There are two ways to form compound words, those are

1. Compounding two or more words together and separating them by a blank space. For example the compound word *rumah sakit* (hospital) that consists of the words *rumah* (house, home) and *sakit* (sick, get hurt, or pain).
2. Compounding two or more words together without separating them by a blank space. In this case the whole compound word is then become a whole new stable term. For example the compound word *kacamata* (eyeglass) that consists of the words *kaca* (glass) and *mata* (eye).

Reduplication – this is an operation that reduplicates the current word form and separating them with a hyphen (except partial reduplication). This operation can be grouped into three different groups and those are

1. Full reduplication, where the reduplication purely duplicates the whole current word form. For example the word *buku-buku* (books) that is derived by repeating the word *buku* (book).
2. Imitative reduplication, where the reduplication of the word form is not identical although they are somehow sounds similar. For example the word *sayur-mayur* (vegetables) that is derived from the word *sayur* (vegetable).
3. Partial reduplication, where the reduplication takes place in the base word itself by placing and additional syllable in the beginning of the word. For example applying the partial reduplication on the word *tua* (old) will derives the word *tetua* (elders).

There are also false reduplications such as in the word *kupu-kupu* (butterfly) which the word itself is already a single unit.

Clitic – is divided into proclitic and enclitic according to its position. One of the clitic examples are clitic that represents independent pronoun, which is grammatically equivalent if replaced by the corresponding pronoun.

Affixation: Clitic	*Word with Clitic*	*Equivalent to*
ku- + *kirim*	*kukirim* (I deliver)	*aku* (I) *kirim* (deliver)
buku + *-nya*	*bukunya* (his/her book)	*buku* (book) *dia* (he/she)

Table 2.7: Clitic Operation Examples

Those four operations can occur as a combination of operations on one single lemma and the features each operation brings are added to set of features of the current word form.

2.2.2 Language Features

Morphosyntactic operations described in the previous chapter bring language features to the affected words or deriving it into new words. Here describe some of the language feature that the operation brings

Gender – There is no explicit and obligatory gender morphological feature, even the pronouns are also not being distinguished by gender, even though we might find suffixes occurs in some rare cases that can distinguish the genders (feminine and masculine). This fashion is rarely used and not productive anymore. The gender classification is not used in the grammar. One of the examples on this gender suffix is the suffix *-wan* and *-wati*.

warta (news) + *-wan* → *wartawan* (news journalist, masculine)
warta (news) + *-wati* → *wartawati* (news journalist, feminine)

Voice – is simply can be identified by the prefix that is added to the Verb. By default a verb lemma is in active voice. To make it explicit or for word derivations purpose, the voices can be marked by prefixes *meN-* and *di-* for active and passive voice respectively.

meN- + *makan* (to eat) → *memakan* (eat)
di- + *makan* (to eat) → *dimakan* (being eaten)

There are also other prefixes such as *ber-* and *ter-* which also marked verb voices. Prefix *ber-* in some cases has brings different feature in both language, for example

MS1:	*Surat*	*itu*	*sudah*	*di+taip*
MS2:	*Surat*	*itu*	*sudah*	*ber+taip*
ID1:	*Surat*	*itu*	*sudah*	*di+ketik*
ID2:	**Surat*	*itu*	*sudah*	*ber+ketik*
	(Letter)	(that)	(already)	(typed)

That letter is already typed

Number – There is no explicit number morphological feature as singular and plural in English, but the reduplication shows plurality of the words (1). Nouns that are preceded by numeral or plural determiner have plural sense but it is not morphologically marked (see number-noun agreement in 2.1). Reduplication that occurs in nouns shows the plurality of the entity of the nouns. For example

REDUP + *buku* (buku, Singular) → *buku-buku* (book, Plural)
REDUP + *sayur* (vegetable, Singular) → *sayur-mayur* (vegetable, Plural)
REDUP + *tamu* (guest, Singular) → *tetamu* (guest, Plural)

Reduplication that occurs on other PoS also shows plurality in another aspect, for example reduplication that occurs on Verb.

REDUP + *aduk* (mix) → *aduk-aduk* (mix repeatedly)
REDUP + *ter* + *jatuh* (fall) → *terjatuh-jatuh* (fall/stumble repeatedly)
REDUP + *kurus* (slim) → *kurus-kurus*
(slim property associated to plural referents)
REDUP + *kurus* (slim) → *kurus-kurus* (rather slim)

Beside the plurality aspect, there is also the prefix *se-* which can be applied on some words that has the meaning of one, same or single.

se- + *nasib* (destiny) → *senasib* (same destiny / single destiny)
se- + *ayah* (father) → *seayah* (same father)
se- + *ekor* (tail) → *seekor* (lit. a tail, the determiner ‘a/an’ for animal)
se- + *gelas* (glass) → *segelas* (one/a glass)

Chapter 3

Machine Translation Platform

Apertium

The practical application of shallow transfer rule-based machine translation method of `id-ms` takes shape by utilizing Apertium (`http://www.apertium.org`), a free/open-source Machine Translation (MT) platform for developing rule-based MT system (2). This platform was initially developed for Romance languages of Spain (3), with Spanish `(es)`, Catalan `(ca)`, and Galician (`gl`) as its main languages at that time. It uses a shallow-transfer word-for-word MT engine that produces fast, reasonably intelligible and easily correctable translations not only for closely related language pairs but also can be extended for language pairs which are not. The free/open-source license encourages the development of new language pairs. Currently there are several language pairs with different typology of languages that are implemented on Apertium. Those languages with different typology also promote new improvement in Apertium. This `id-ms` language pair implementation on Apertium for example promotes the integration of Foma (`http://foma.sourceforge.net`), a finite-state toolkit (4), to Apertium which will be discussed in more detail on Chapter 4.

Foma

3.1 Architecture

Apertium has a modular architecture consists of these following modules

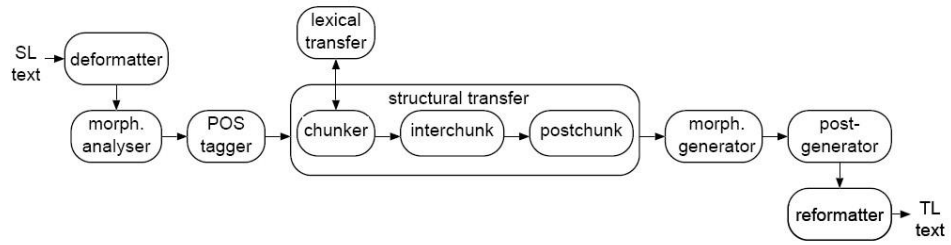


Figure 3-1: The modular architecture of Apertium

The pipeline of modules starts with the *deformatter* module which encapsulates the input document formatting information. This module will make the document seen as cleaned text of words separated with blank space before passing it to the rest of the modules. And as a mirror, the *reformatter* at the end of the pipeline recovers the document formatting information that was encapsulated by the *deformatter*. This is

simply the opposite direction of the *deformatter* module. Since these two modules are not the core of the work, these two modules will not be explained in more detail.

There is another module that is not used in `id-ms` new MT system, which is the *post-generator* module. This module performs orthographic operations of contractions and apostrophations which occur in Romance languages. This module is not used in `id-ms` MT system since the operations are not occurs in those languages.

Then the focus of the development starts from `id` morphological analyzer through `ms` morphological generator. The descriptions of those modules are as the following:

- A *morphological analyzer* tokenizes the lexical units which are the surface form words in the text. Then it produces the analysis for each of the lexical unit in the form of *lemma* followed by *lexical category* tag and *morphological inflection information* tags. For example, the `id` sentence *"email itu sudah dapat kuketik"* ("The email can already be typed by me") is analyzed into

```
^email/email<n><bare><sg>$
^itu/itu<det>$
^sudah/sudah<adv>$
^bisa/bisa<mod>/bisa<n><bare><sg>$
^kuketik/aku<prn><p1><sg><pro>+ketik<vblex><pasv><imp><sg>$
```

- *PoS tagger* is a first-order (bigram) HMM PoS tagger using supervised or unsupervised classical method (Baum-Welch algorithm) to estimate its parameters by training it on monolingual source language (SL) corpora or a novel unsupervised approach utilizing the MT engine and target language (TL) corpora (5). This tagger disambiguates the ambiguous analysis, as for the example above the PoS tagger disambiguates the word *bisa* as a modal verb 'can' instead of a noun which in that case means 'snake venom'.

```
^email/email<n><bare><sg>$
^itu/itu<det>$
^sudah/sudah<adv>$
^bisa/bisa<mod>$
^kuketik/aku<prn><p1><sg><pro>+ketik<vblex><pasv><imp><sg>$
```

- The *lexical transfer* delivers the TL lexical form of the corresponding SL lexical form by looking up a bilingual dictionary. The dictionary is in XML format as below

```
<e><p><l>email<s n="n"/></l>
       <r>emel<s n="n"/></r></p></e>
<e><p><l>itu<s n="det"/></l>
       <r>itu<s n="det"/></r></p></e>
<e><p><l>sudah<s n="adv"/></l>
       <r>sudah<s n="adv"/></r></p></e>
<e><p><l>bisa<s n="mod"/></l>
       <r>boleh<s n="mod"/></r></p></e>
<e><p><l>aku<s n="prn"/></l>
       <r>aku<s n="prn"/></r></p></e>
<e><p><l>ketik<s n="vblex"/></l>
       <r>taip<s n="vblex"/></r></p></e>
```

- A *structural transfer* consists of three sub-modules, *chunker, interchunk,* and *postchunk*. This `id-ms` MT system only uses one of the sub-modules, *chunker,* which invokes the *lexical transfer* and performs local syntactic operation according

to hand-made rules written in XML. For example this id-ms MT system prefers to change the clitic of the words into independent word.

```
kuketik (id) → kutaip (ms) → aku taip (ms)
```

The structural transfer produces an analysis in TL side.

```
^emel<n><bare><sg>$
^itu<det>$
^sudah<adv>$
^boleh<mod>$
^aku<prn><pl><sg>$
^taip<vblex><pasv><imp><sg>$
```

- The *morphological generator* generates the surface form of the words in TL side given the analysis. This is simply the opposite direction of the *morphological analyzer* module. As for example above, it produces the surface form of the TL form of the given analysis as

```
emel itu sudah boleh aku taip
```

The detailed development of each module will be discussed in the next chapters.

3.2 Resources

This platform encourages to record linguistic data of the language pairs mostly in XML format. Developing id-ms MT system includes adapting linguistic resources of those languages into Apertium environment, such as the finite-state morphology to do analysis and generation including the monolingual dictionaries, HMM Pos Tagger, bilingual dictionary, and the transfer rules.

The language resources of Indonesian and Malaysian are not fine-grained yet which make us decided to build most of the language resources for the modules from scratch to make this id-ms MT system running. The id-ms MT system is designed to make the later improvement on the language resources manageable and easy.

The development of the finite-state morphologies for id-ms pair is using a different technology compare to most of those Apertium stable language pairs since id and ms have different kind of morphologies phenomena. This is also because the morphological analyzer modules of Apertium were initially designed to handle Romance language morphologies, which are different to id or ms. The id-ms morphologies are handled by a different tool that is not the default of Apertium and has not been used in any other stable language pairs before. This promotes integration of a new tool to Apertium platform with their new added-values and new challenges. By using this different tool, the format of the monolingual dictionary is not encoded in XML format as it generally was in the current stable language pairs. The detailed description of the development including the tools selection will be discussed in Chapter 4 Finite-State Morphology.

During the MT development, id-ms MT is reusing the other language pair HMM PoS tagger, since this is not a crucial for the MT development at the moment. The creations of the monolingual dictionaries and bilingual dictionary will be discussed in Chapter 5. The transfer rules and several additional processing will be discussed in Chapter 6.

Chapter 4

Finite-State Morphology

The engineering task of the `id-ms` MT system mainly falls on building the finite-state morphology tool. Since both languages under question have similar morphological mechanism, the tools could be easily adapted to each other. The finite-state morphology divided into morphological analyzer and generator tool. The morphological analyzer produces the analysis for the source language (SL) side, `id`, while the morphological generator generates the surface form of the target language (TL), `ms`, given its analysis after transfer takes place.

Since the `id-ms` pair has different morphology to the Romance languages, the finite-state morphology tools for `id-ms` pair is not using the default tool provided by Apertium. This chapter describes the development of the finite-state morphology including the tool selection and then followed by the development of the monolingual dictionary which has different format than the other Apertium stable pair monolingual dictionaries.

4.1 Tool Selection

The morphological analyzer will convert word in surface form into their lemma and followed by the morphological tags, which are consists of *lexical category* tag and *morphological inflection* tags. The *morphological inflection* tags correspond to morphosyntactic operations of the word. For example the analysis of an Indonesian word *tercantik* (most beautiful) will be *cantik*`<adj><sup>`, where it consists of the lemma *cantik* (beautiful), the *lexical category* tag `<adj>` which represents adjective, and the *morphological inflection* tag `<sup>` which represents superlative features. The morphological generator simply performs the opposite direction.

The initial morphological tool that is widely used by Apertium stable language pairs is *Lttoolbox*, a toolbox for lexical processing, morphological analysis and generation of words, which has been used on several language pairs, among many others are Spanish-Catalan (`es-ca`), Spanish-Portuguese (`es-pt`), Swedish-Danish (`sv-da`). Mostly those language pairs are languages that only have morphological operations occurring in the end part of the word which Apertium was initially designed for.

Lttoolbox

Lttoolbox works by listing exhaustive combination of the inflections which occur in a language, called paradigm. This tool has several limitations on accommodating the language pairs under questions, Indonesian and Malaysian. Those limitations are

- *Morphemes that precede the base word are not handled straightforwardly*. In the case of Indonesian and Malaysian, there are numerous morphemes preceding the base form. Unfortunately in Lttoolbox, the process of the analysis takes place on the morpheme's position. Therefore the prefix analysis, which is the tag(s), will be in the front of the lemma. Therefore, a separate additional post-formatting for the tags positioning needs to be done

- *Noncontinuous morphemes are handled independently*. In the case circumfix, the morphemes that compose the circumfix will be treated as independent prefix and suffix.

- *Each morphophonemic case is treated as one independent paradigm.* The morphophonemic rules are handled by expanding the morpheme to its all possible forms. For example the prefix *meN-* will be expanded to its different forms regarding to which base word it glued i.e. it will change into *menge-* for one syllable case, *meng-* for base words starting with [a i u e o g h], *meny-* for base words starting with [s, y] and so on which all those will be treated as different paradigms.

- *This tool cannot handle reduplication cases.*

By these arguments it is decided not to use *Lttoolbox* and employs an available Indonesian morphological analyzer (6), hereinafter will be called the initial morphological analyzer tool. This initial tool was developed on XFST (7), Xerox finite state tools, and LEXC, high-level declarative language to specify language lexicon. This tool covers the reduplication case and the morphophonemic rules and it also includes an Indonesian monolingual dictionary composed by a large number of Indonesian lemmata. Unfortunately the coverage of how it handles the morphosyntactic operations is not adequate enough for the task, where

- *It covers partly the morphological operations.* It only handles several morphological operations such as reduplication and partly affixations cases, not including clitics and particles. The uncovered cases will cause the inflected word to be left un-translated.

- *The tagset is underspecified for the generation.* The tagset composed of 17 general tags including the Part-of-Speech (PoS) tags and some of the morphological operations that occur (see Table 4.1). The PoS tags simply mark three PoS types, namely Verb, Noun, and Adjective, while others are considered as Etc.

Tag	*Description*
`+Verb`	Verb tag
`+Noun`	Noun tag
`+Adjective`	Adjective tag
`+BareVerb`	Verb tag in lemma form
`+BareNoun`	Noun tag in lemma form
`+BareAdjective`	Adjective tag in lemma form
`+BareEtc`	Tag for any undefined lemmata
`+AV`	Tag for active voice
`+PASS`	Tag for passive voice

+UV	Tag for undergoer voice
+Redup	Tag marking the reduplication
+Caus_kan	Tag marking the suffix -kan, that is form of causative that needs an actor for the event
+Appl_kan	Tag marking the suffix -kan, that is form of causative that does not need an actor for the event
+Caus_kan	Tag marking the suffix -i, that is form of causative that needs an actor for the event
+Appl_kan	Tag marking the suffix -i, that is form of causative that does not need an actor for the event
+Actor	Tag for actor
+Instrument	Tag for instrument

Table 4.1: Initial Morphological Tool Tagset

- S*everal inflected words have the same analysis*. This is unfavorable for the translation task since those different inflected words will be transferred to the same target analysis. For example in the case of a verb derived noun, *kiriman* (package), *pengirim* (deliverer), and *pengiriman* (delivery) from the verb *kirim* (to deliver), those will have the same `kirim+Noun` as the analysis.

- Still referring to the analysis problem, *the generation step generates a big number of different inflected words for the same analysis*, which will produce bigger numbers of translation hypotheses.

Analysis			*Generation*		
kiriman *pengirim* *pengiriman*	>	`kirim+Noun`	`kirim+Noun`	>	*pengirim* *pengiriman* **pemberkiriman* **perkiriman* **kepengiriman* **keberkiriman* **kekiriman* *kiriman*

Table 4.2: Initial Morphological Analysis and Generation.
(*) marks a non valid inflected words

Although the initial tool is not adequate for the task, it is a good start point for the development of the morphology module. Starting off from that, a new morphological analyzer was developed using the same technology that is XFST and LEXC (xfst/lexc), which based on finite-state transducer technology. The integration of the tool to the module in Apertium is done by using *Foma*. The integration of xfst/lexc through Foma is relatively new to Apertium, which is then found several cases than need special treatments (for detail see Chapter 6).

4.2 Tagset

The new morphological analyzer/generator, hereinafter will be referred as morphological analyzer, composed with new finite state structures to cover more morphological operations and we also re-used the morphophonemic rules and reduplication feature from the initial tool. The morphosyntactic operations that occur in both languages are assumed to be similar. By this both of the languagees sides are using the same finite state structures with different lexical entries.

We changed the form of the tag from `+TAG` into `<TAG>` to suit the Apertium platform. The new morphological analyzer uses a new fine-grained tagset to encounter the underspecified analysis and some of tags are tags that is commonly use in Apertium. Given in Table 4.4 the tagset that is used in the system.

Comparing to the previous example in Table 4.2, with the morphological analyzer the analysis are more precise as seen on the Table 4.3 below

Analysis		
kiriman	>	`kirim<vblex><ent><sg>`
pengirim	>	`kirim<vblex><actor><sg>`
pengiriman	>	`kirim<vblex><actio><sg>`

Generation		
`kirim<vblex><actor><sg>`	>	*pengirim*
`kirim<vblex><actio><sg>`	>	*pengiriman*
		**#pemberkiriman*
		**#perkiriman*
		**#kepengiriman*
		**#keberkiriman*
		**#kekiriman*
`kirim<vblex><ent><sg>`	>	*kiriman*

Table 4.3: Morphological Analysis and Generation.
(*) marks a non valid inflected words
(#) marks the un-generated inflected words

The morphological tags that follow the lemma have fixed order as shown on the scheme below. The square bracketed tags are tags that occur on special cases such as to mark capitalized words.

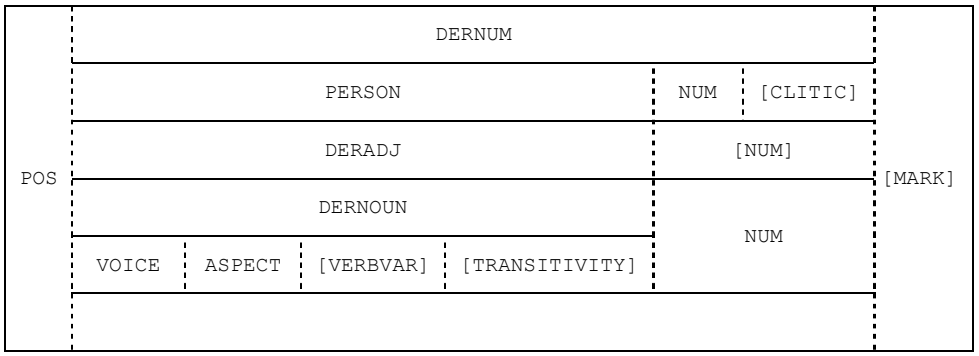


Figure 4-1: Morphological tags fixed order schema

Tag	*Description*	*Tag Type*
<adj>	adjective lemma	POS
<n>	noun lemma	POS
<num>	number lemma	POS
<prn>	pronoun	POS
<det>	determiner	POS
<cnjcoo>	coordinating conjunction	POS
<cnjsub>	subordinating conjunction	POS
<vblex>	verb lemma	POS
<part>	particle	POS
<mod>	modal	POS
<ij>	interjection	POS
<qst>	question word	POS
<pr>	preposition lemma	POS
<p1>	first person	PERSON
<p2>	second person	PERSON
<p3>	third person	PERSON
<sg>	singular	NUM
<pl>	plural	NUM
<card>	cardinal number	DERNUM
<ord>	ordinal number	DERNUM
<coll>	collective number	DERNUM
<ref>	referential number	DERNUM
<vbhaver>	verb ‘to have’	VERBVAR
<vbser>	verb ‘to be’	VERBVAR
<actv>	active voice	VOICE
<pasv>	passive voice	VOICE
<perf>	perfective aspect	ASPECT
<imp>	imperfective aspect	ASPECT
<bare>	bare noun	DERNOUN
<abstract>	derived abstract noun	DERNOUN
<actio>	derived action noun	DERNOUN
<actor>	derived actor noun	DERNOUN
<ent>	derived entity noun	DERNOUN
<theme>	derived theme noun	DERNOUN
<positive>	bare adjective	DERADJ
<sup>	superlative adjective	DERADJ
<exceed>	adjective that shows something exceeding	DERADJ
<manner>	adjective that shows similar manner	DERADJ
<uni>	union adjective	DERADJ
<possib>	adjectival phrase	DERADJ
<mimic>	mimicking manner adjective	DERADJ
<enc>	enclitic	CLITIC
<pro>	proclitic	CLITIC
<appl>	applicative	TRANSITIVITY
<caus>	causative	TRANSITIVITY
<cap>	capitalization mark	MARK
<pos>	possesive mark	MARK

Table 4.4: Morphological Analyzer Tagset

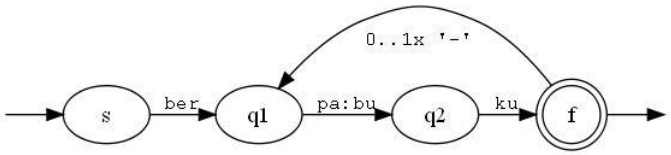
The complete tagset list and the example of the tagset usage can also be found at the wiki page of Apertium for Indonesian and Malaysian language pair
http://wiki.apertium.org/wiki/Indonesian_and_Malaysian#Tagset

4.3 Basic notation

The design structure of the finite state morphology is represented in form of *transition diagram* of a *Finite State Transducer* (FST) with additional information on the edges for reduplication purpose. The construction of the FST is a 6-tuple (Q, Σ, Γ, I, F, δ), where,

- Q is a set of *states*;
- Σ is a set of *input alphabet*;
- Γ is a set of *output alphabet*;
- I is a subset of Q, as start states;
- F is a subset of Q, as final states; and
- $\delta \subseteq Q \times (\Sigma \cup \{\varepsilon\}) \times (\Gamma \cup \{\varepsilon\}) \times Q$ (where ε is an *empty string*), is the *transition relation*.

The additional information defines the number of times to pass a transition. Here is an example of a simple FST *transition diagram* structure with the additional information:



Where *s* is the start state and *f* is the final or accepting state. The transition from *q1* to *q2* has a rewriting rule that changes the string *'bu'* to *'pa'*, while in transitions *s* to *q1* and *q2* to *f* the rewriting rule is shortened and just in the form of *'ber'* and *'ku'* while in fact it changes the string to the same string, explicitly can be written as '*ber:ber*' and *'ku:ku'*. The additional information is on the transition edge between *f* and *q1*, where it states that it has to pass zero to one time through that edge with an added '-' (hyphen) character. This structure of *transition diagram* rewrites the string *'berbuku'* into *'berpaku'* and *'berbuku-buku'* into '*berpaku-paku*'.

4.4 Finite state structure

The overall finite-state structure of the `id-ms` system can be found in Figure 4-2. There is a corresponding finite-state structure for each morphosyntactic rule. Given in Figure 4-3 is the example of several numeral alternations which recognizes the cardinal, ordinal, and collective numerals. For example this structure with lemma *tujuh* (seven) as one of the Numeral entries will recognize the cardinal form *tujuh* (seven), the ordinal form *ketujuh* (seventh), and collective form *bertujuh* (in group of seven).

Flag Diacritics

We use xfst Flag Diacritics in the finite state structure. The Flag Diacritics are just a normal multicharacter that can be treated as regular string but they have special form, which is in the form of @FLAG.FEATURE.VALUE@, and treated specially by xfst routines. The flag diacritics here will be used to signal which lexical category tags and morphological information tags that should be in the analysis and also for restricting the finite state path. The replacement of the triggered flag diacritics to tags is done in the *tagging* node. The complete finite state structure can be found in Appendix A Finite-State

Transducers Structure. The descriptions of the diacritics flags that are used are available in Appendix B Flag Diacritics List.

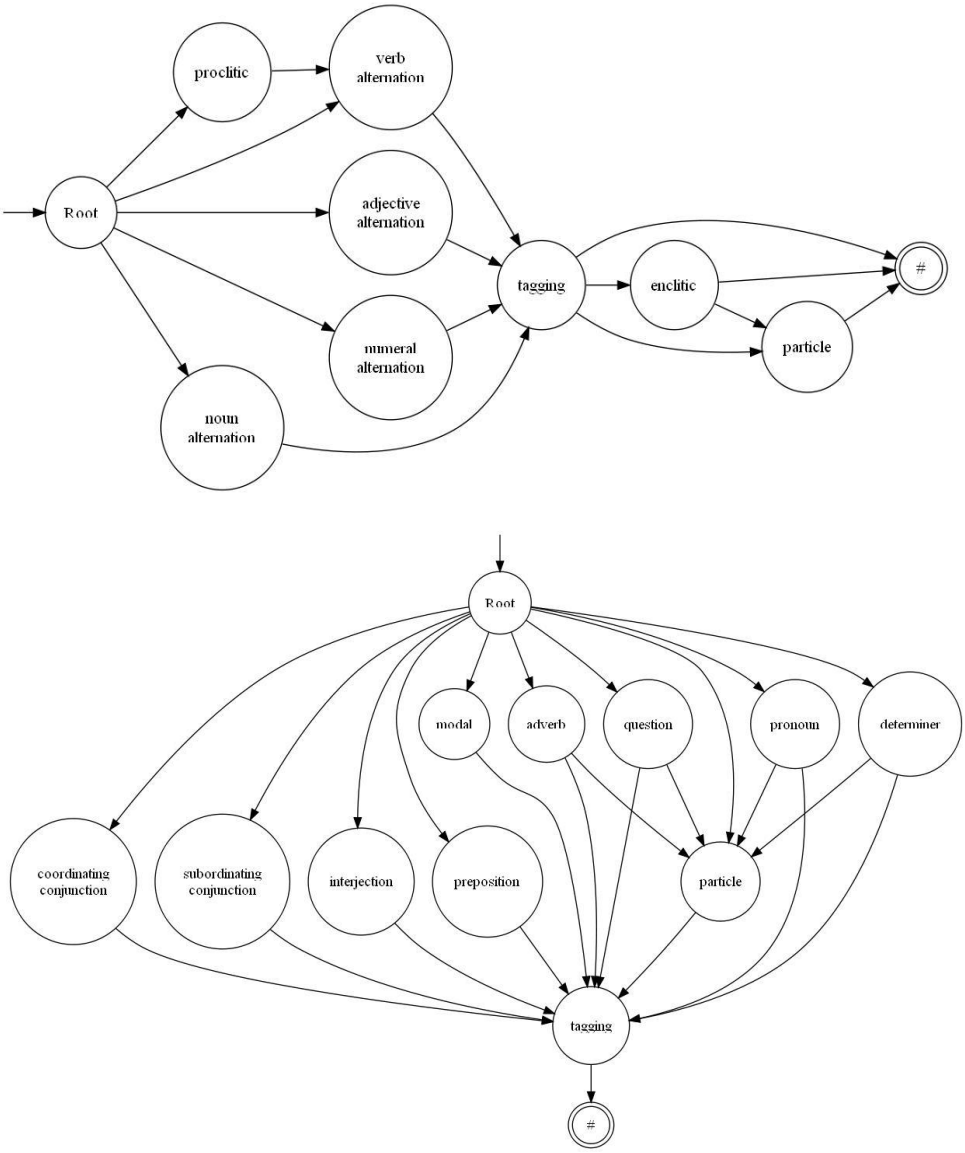


Figure 4-2: Overall id-ms MT finite-state transducer structure

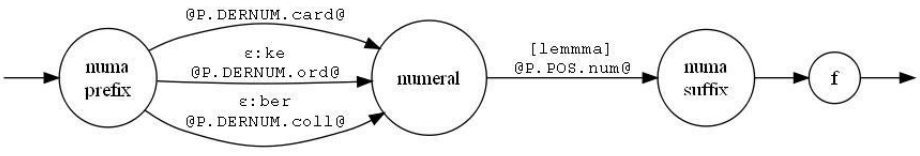


Figure 4-3: Finite-state structure example for several numeral alternations (cardinal, numeral, and collective numerals)

4.5 Compilation

Some parts of the code in the initial tool that handle the reduplication is patent-encumbered, which is not preferable to Apertium since it has to keep its free/open-source license. In order to that the Xerox function `compile-replace()` is replaced into sequence of several functions, which gives the same outcome. The replacement gives some minor problem on the compilation in Foma.

The reduplication is done by copying the part which marked as to be reduplicated. All the copies are paired to one another and then the pairs which are not similar are being sorted out. This algorithm needs a greater memory and takes longer time during the compilation especially with bigger number of lexical entries.

To encounter that problem we try to reduce the search space of the non-similar pairs. Since the lexical entries of the monolingual dictionaries are coded together with the morphological analyzers, we divide the lexical entries into few parts of smaller number of entries. Then we compile it separately and combined them together.

The division is done automatically by a script which takes the core finite state part of the morphological analyzer as a template and a complete list of lexical entries. The script will divide the lexical entries and put it together it with the morphological parts template file, compiles them separately and at the end combines them together. This is also makes a clear division between both parts and makes the future changes in lexical entries more manageable.

The lexical entries list file is recorded in this format below, where the lexical categories are in the left part of the '`||`' and the actual code of the lexical entries are on the right side.

```
noun||email@P.POS.n@                    branching ;
noun||tamu@P.POS.n@@P.NUM.pl@           branching ;
noun||tetamu@P.POS.n@@P.NUM.pl@         branching ;
...
verb||ketik@P.POS.v@                    branching ;
verb||kirim@P.POS.v@                    branching ;
...
```

There are 16 lexical categories nodes including `exception` that is provided if there are some rare cases that need special treatment. Those 16 lexical entries are

- `noun`
- `verb`
- `adjective`
- `adverb`
- `coordinating_conjunction`
- `subordinating_conjunction`
- `particle`
- `modal`
- `numeral`
- `pronoun`
- `preposition`
- `question`
- `interjection`
- `determiner`
- `punctuation`
- `exception`

The flag diacritic that signals morphological tags such as singular of plural can also be added in the lexical entries.

Chapter 5

Monolingual and Bilingual Dictionaries

Since the computational linguistics researches on id and ms are not as enthusiastic as to compare to European languages, it is difficult to find language resources to support the system. The creation of id and ms monolingual dictionaries and id-ms bilingual dictionary is one of the attempts for this id-ms MT system to have language resources in a reasonable prototype quantity and quality. From the design perspective, the idea is to have dictionaries that are easy to manage for future correction or additional lexical entries.

monolingual dictionary

5.1 Comparable Corpus

The approach which was taken to build the dictionaries starts by making a small size comparable corpus which will be the base for the lexical entries of the dictionary and mapping of words between id and ms. By this, the size of the dictionaries are relatively small but having a reasonable quantity and quality on an actual comparable sentences.

The comparable corpus composed of sentences from manuals of home appliances (1138 sentences), product license agreements (204 sentences), and holy books (3832 sentences). It consists of 5,174 sentences in total for each language side.

The sentences in both languages are aligned and being checked manually, since the size is relatively small. Based on human judgment, the pairs of sentences that came from manuals and license agreement are having better word-to-word mapping than holy books which in holy books the sentences have different way of paraphrasing because of the cultural difference. This causes some problem on the creation of the bilingual dictionaries, since this source is the biggest part of the comparable corpus.

5.2 `ID` Monolingual Dictionary

Indonesian monolingual dictionary

As the starting point to build the `id` monolingual dictionary is the lexical entries which are list of lemmata provided in the initial `id` morphological analyzer. This initial tool includes lexical entries tagged as 'verb', 'adjective', 'noun', and 'etc'. Further in detail it consists of

- 3,417 verb tagged lexical entries,
- 19,036 noun tagged lexical entries,
- 5,863 adjective tagged lexical entries, and
- 4,153 etc tagged lexical entries

Figure 5-1: Initial ID-morphological analyzer lexical categories with its corresponding number of lexical entries

Several of those tagged lexical entries are found to be tagged incorrectly across the lexical categories and moreover the lexical entries which are tagged as 'etc' are underspecified.

Those lexical entries then chosen and adapted to fit the `id-ms` MT system design. The `id-ms` MT system considers 14 lexical categories and those categories with its corresponding number lexical entries can be found in Figure 5-2.

The closed word class lexical categories such as preposition or particle are added manually. As for noun, adjective, and verb, the lexical entries are taken from the initial tool, handpicked to fit the `id` sentences of the comparable corpus; this is done to retain the translation quality. Since the work is a manual work, instead of having a large number of lexical entries as in the initial tool has, the `id-ms` MT system consists of

- 1,300 noun lexical entries
- 459 verb lexical entries
- 438 adjective lexical entries
- 18 numeral lexical entries
- 13 pronoun lexical entries
- 92 adverb lexical entries
- 15 preposition lexical entries
- 4 particle lexical entries
- 3 coordinating conjunction lexical entries
- 30 subordinating conjunction lexical entries
- 19 determiner lexical entries
- 9 interjection lexical entries
- 7 question lexical entries
- 4 modal lexical entries

Figure 5-2: Current ID-morphological analyzer lexical categories with its corresponding number of lexical entries

An evaluation is performed to make a comparison of the current `id` morphological analyzer with the initial morphological analyzer in terms of inflection and lexical entries coverage. The evaluation is done by taking random 1000 sentences of Indonesian text from Wikipedia dump[1]. Then the text is analyzed by both morphological analyzers.

The metrics used in the evaluation is the same metric that is used to measure the language resources of stable Apertium language pairs (8) although in this evaluation the corpus is relatively small. The surface words that are being considered in this evaluation are only words containing alphabet and hyphen, and not digits.

[1] http://download.wikipedia.org/ - Wikipedia dump of 19th June 2008

morphological analyzer	*lemmata*	*surface*	*mean ambiguity*	*coverage*	*corpus*
`ID`-Initial Morphological Analyzer	32,469	14,410 (15,897)	1.33	66.9%	1000 sentences taken randomly from Wikipedia
`ID`-Current Morphological Analyzer	2.411		1.19	54.88%	

Table 5.1: Initial and Current ID-Morphological Analyzer analysis comparison

The *surface* is the number of surface forms recognized by the morphological analyzer where compound words are disregarded; the *mean ambiguity* is the number of lexical analyses per surface form; the *coverage* is a naïve measurement of surface forms fraction for which returns at least one analysis.

Although there was big number of reduction that has been made on the *lemmata* column for the new `id` morphological analyzer, it only brings a slight difference on the *coverage* column. And the current `id` morphological analyzer is more precise as shown on *mean ambiguity* column. The *coverage* of the current `id` morphological analyzer can be improved by adding more correctly tagged lemmata.

Another thing that can be evaluated is the ambiguity that both morphological analyzers have on the generation phase. Here is the evaluation summary with the same settings but with different direction

morphological analyzer	*surface*	*analyzed*	*#analysis*	*#generation*	*mean generation*
`ID`-Initial Morphological Analyzer	14,410	9640	12,817	34,971	2.73
ID-Current Morphological Analyzer		7909	9.375	11,791	1.26

Table 5.2: Initial and Current ID-Morphological Analyzer generation comparison

The *analyzed* is the number of surface words that has at least one analysis; the *#analysis* is the total number of analysis that the previous evaluation produced; the *#generation* is the total number of generation of those analysis; the *mean generation* is the fraction of those last two numbers. As seen on Table 5.2, the generation process of the current `id` morphological analyzer is less ambiguous than the initial tool.

Provided below an example on the evaluation calculation of the surface words "*Saya menulis email*" (I write email).

Input text: "Saya menulis email"

Initial id morphological analyzer

*Saya	-
tulis+Verb+AV	*membertuliskan *mempertuliskan *menertuliskan menuliskan menulisi menulis
email+BareNoun	email

Current id morphological analyzer

saya<prn><p1><sg><cap>	Saya
tulis<vblex><actv><imp> <sg>	menulis
email<n><bare><sg>	email

morphological analyzer	*surface*	*analyzed*	*#analysis*	*#generation*	*mean ambiguity/ generation*	*coverage*
ID-Initial Morphological Analyzer	3	2	2	7	1 / 3.5	66.67%
ID-Current Morphological Analyzer		3	3	3	1 / 2	100%

Figure 5-3: Morphological analyzer resource evaluation calculation examples

Surface word examples on current id morphological analyzer above are words with only one generation for each analysis. On the current tool, when there are more generations for each analysis, mostly because it generates the synonyms. For example the word '*jika*' (if), it will produce 3 output if it is regenerated, which all of them are synonyms.

jika → apabila<cnjsub> → *apabila*, *jika*, *bilamana*

The monolingual dictionary entries are part of the xfst/lexc morphology analyzer source code. The entries are divided to each lexical category so that the future improvement for the monolingual dictionary is easy to manage. Below is the example of the lexical entries for the input sentence above when they are put together with the morphological operation finite state code.

```
LEXICON pronoun
saya@P.POS.prn@@P.PERSON.p1@@P.NUM.sg@          branching ;

LEXICON verb
tulis@P.POS.v@                                  branching ;

LEXICON noun
email@P.POS.n@                                  branching ;

LEXICON subordinating_conjunction
apabila@P.POS.cnjsub@:jika000@P.POS.cnjsub@     branching ;
```

```
apabila0@P.POS.cnjsub@:bilamana@P.POS.cnjsub@   branching ;
apabila@P.POS.cnjsub@                           branching ;
```

Or as shown below if they are put in separate file to handle the compilation minor problem as mentioned in subchapter 4.5.

```
pronoun||saya@P.POS.prn@@P.PERSON.p1@@P.NUM.sg@       branching ;
verb||tulis@P.POS.v@                                  branching ;
noun||email@P.POS.n@                                  branching ;
subordinating_conjunction||apabila@P.POS.cnjsub@:jika000
   @P.POS.cnjsub@                                     branching ;
subordinating_conjunction||apabila0@P.POS.cnjsub@:bilamana
   @P.POS.cnjsub@                                     branching ;
subordinating_conjunction||apabila@P.POS.cnjsub@      branching ;
```

5.3 MS Monolingual and Bilingual Dictionary

Not like the id monolingual dictionary creation, currently there is no available ms wordlist and even more the one with lexical category tag information. The attempt to build the dictionary is done in three steps,

- Collect lemmata forms by picking several lemmata form hypotheses of the inflected ms words that occur in the comparable corpus.
- Tag the lexical category of the lemmata by the same tag as their corresponding id word mapping.
- Handpicked the result.

Malaysian monolingual dictionary

By this, the creation of the ms monolingual dictionary is done in parallel with the creation of id-ms bilingual dictionary.

The first step is done by making a list of surface words of the ms side of the comparable corpus which most of them are inflected words. Then the morphemes of those surface words are stripped to get their lemma hypotheses. This is done by creating a Finite-State Transducer (FST) that will strip the morphemes of the inflected verb, adjective, and noun which then returns several hypotheses. For example running the word *menghantarkan* on that FST which has the correct lemma form *hantar* will return

```
^menghantarkan/menghantarkan<n><bare><sg>/menghantark
<adj><ent><sg>/menghantark<vblex><ent><sg>/menghantar
k<n><ent><sg>/menghantarkan<adj><positive>/ …
hantar<adj><actv><imp><caus><sg>/hantar<vblex><actv><
imp><caus><sg>/hantar<n><actv><imp><caus><sg>/ …
```

Those all hypotheses are disambiguated by a simple voting only on the lemma (without taking into account the morphological tags) of several inflected word that is in the same group. By group it means that those inflected words that at least having one similar hypothesis. For example the words *menghantarkan*, *dihantarkan*, *hantarkanlah*, *hantar*, etc, will be in the same group because they share several hypotheses. By this the voting will return *hantar* simply because all of them will have this hypothesis which by voting will stands out among the others. The most probable lemma hypotheses will be mapped and recorded to its inflected words.

The second step is to get the lexical category of those ms lemmata hypotheses. The assumption is that the ms lemmata will have the same lexical categories as their corresponding id lemmata. The mapping is done by training the comparable corpora using Moses (9) and takes the translation table output. The source languages of the training are the analysis of id sentences by using the id morphological analyzer. The output on the translation table is as given below

```
^kirim<vblex><av><perf><caus><sg><par>$ hantarkanlah 1.0000
...
^tuang<vblex><av><perf><caus><sg><par>$ curahkanlah 1.0000
...
^kabel<n><bare><sg>$ kuasa 0.0076336
^kabel<n><bare><sg>$ kord 0.0909091
```

Then we make a temporary ms morphological analyzer. The lexical entries for the ms temporary morphological analyzer are the lemmata hypotheses of the mapping above, those are

- *hantar* for *hantarkanlah*,
- *curah* for *curahkanlah*,
- *kuasa* for *kuasa*, and
- *kord* for *kord*

And those lemmata hypotheses are categorized as their corresponding id lexical category tags. In this example will be the lemma *hantar* and *curah* as verbs and *kuasa* and *kord* as nouns. Then we do once again but with both side analyzed. The output on the translation table is as given below

```
^kirim<vblex><av><perf><caus><sg><par>$
   ^hantar<vblex><av><perf><caus><sg><par>$ 1.0000
...
^tuang<vblex><av><perf><caus><sg><par>$
   ^curah<vblex><av><perf><caus><sg><par>$ 1.0000
...
^kabel<n><bare><sg>$ ^kuasa<n><bare><sg>$ 1.0000
^kabel<n><bare><sg>$ ^kord<n><bare><sg>$ 1.0000
```

The most probable correct pairs for the bilingual dictionary are picked by measuring the edit distance of the tags of the candidate pairs with Levenshtein algorithm by also taking account the probability that the translation table has given. The score is the probability given in the translation table divided with the Levenshtein distance. Only pairs that have the best score are taken as entry in bilingual dictionary.

The result of this process is also briefly checked manually to resolve the morphophonemic ambiguity that is still preserved.

At the end of this process, the bilingual dictionary consists of 2395 entries and the ms monolingual dictionary is created by taking the lexical entries of the ms side of the bilingual dictionary.

The entry format of ms monolingual dictionary will be in the same as forms as the id monolingual dictionary. While the entries of the bilingual dictionary is in XML format is as shown below

```
<e><p> <l>bahwa<s n="cnjsub"/></l>
       <r>bahawa<s n="cnjsub"/></r></p></e>
<e><p> <l>enggan<s n="adj"/></l>
```

```
       <r>enggan<s n="adj"/></r></p></e>
<e><p> <l>untuk<s n="pr"/></l>
       <r>untuk<s n="pr"/></r></p></e>
<e><p> <l>setuju<s n="vblex"/></l>
       <r>bersetuju<s n="vblex"/></r></p></e>
<e><p> <l>tentu<s n="adj"/><s n="abstract"/><s n="pl"/></l>
       <r>terma<s n="n"/><s n="bare"/><s n="pl"/></r></p></e>
```

Although some of the lexical entries in the bilingual dictionary are really well mapped, there are also a lot of entries that are not mapped correctly. This is because most of the sentences that come from holy books in the comparable corpus are not word-to-word translation but more towards paraphrasing. So to encounter this it needs a lot of hand-on treatment to pick the correct mapping while there is not enough human resources.

Chapter 6

Transfer and Additional Process

This chapter explains the transfer process in the structural transfer module and the additional translation process.

The structural transfer module invokes the lexical transfer that utilizes the bilingual dictionary. This is not the most crucial part of the work since the structure between the languages is similar and mostly the difference is in the writing style which is influenced by the cultural differences.

Another thing that is explained in this chapter is the additional processes that are added between several the modules to cover several features that are not yet covered by the current platform. The additional process is a temporary tweak while the Apertium community is in parallel working on the permanent solution.

6.1 Transfer

The language pair is closely related grammatically and syntactically which should simplify the process of the translation and assumed not to need a structural transfer phase. But then there are several transfer rules which are defined to reduce the number of translation hypotheses. The reduction happens when there are more hypotheses which are equally equivalent to each other. The method is simply choosing one preferable structure among those that are equivalent.

Should we take a look on an example, such as in the case of clitics where the surface words containing clitics have another surface forms that are equally grammatical. This several forms will be considered as different hypotheses which actually they are grammatically correct and brings the same information. The transfer rule will make the equally grammatical forms to change into only one chosen form, in this case. For example the surface words *kumengirimnya* (I send him/her) will have several other equally grammatical forms which those are *'aku mengirimnya'*, '*aku mengirim dia*', and '*kumengirim dia'* which all words or phrase forms have the same meaning. For this case where the clitics can stands as an independent pronoun, there is a transfer rule to change those surface forms with clitics to be explicitly specifying its clitics as pronouns.

The transfer rules are written in XML format. The following is one transfer rule example to change the proclitic to be an independent pronoun

```
<rule>
   <pattern>
      <pattern-item n="pronpro"/>
      <pattern-item n="verb"/>
   </pattern>
   <action>
      <out>
         <mlu>
            <lu>
               <clip pos="1" side="tl" part="lemh"/>
               <clip pos="1" side="tl" part="a_pos"/>
               <clip pos="1" side="tl" part="a_person"/>
               <clip pos="1" side="tl" part="a_num"/>
            </lu>
            <b/>
            <lu>
               <clip pos="2" side="tl" part="lemh"/>
               <clip pos="2" side="tl" part="a_pos"/>
               <clip pos="2" side="tl" part="a_voice"/>
               <clip pos="2" side="tl" part="a_aspect"/>
               <clip pos="2" side="tl" part="a_num"/>
            </lu>
         </mlu>
      </out>
   </action>
</rule>
```

6.2 Additional Process

Since the integration of Foma to Apertium is very recent and still on-going where this language pair also promotes the process, there are several additional processing that are needed between the modules. Foma as the morphological analyzers compiler compiles each word one by one separately with blank space as the separator. In the current integration, Foma cannot capture the compound words since their individual component are separated with blank space, different than compound words in German. Because of this, additional scripts are added before the analysis and after the generation takes place.

The script before the analysis takes place will match the compound words that are predefined in a compound word list with the surface words in the source sentence given. The match is in the longest match fashion.

If there is a match it will replace the blank space character of the matched compound words into ~ (tilde) character. This will make Foma treat it as one surface word. By this the compound word entries in the monolingual dictionary and bilingual dictionary will be adapted to this form as well. The script after the generation will simply do the opposite direction where it changes the tilde back to blank space character. Here is an example of the translation of the compound words for the word *ibu kota negara* (capital city of a country).

The additional script will change the blank space in the surface form

```
Ibu kota negara -> Ibu~kota~negara
```

Entry in `id` monolingual dictionary the entries are shown as the following

```
LEXICON noun
ibu%~kota%~negara@P.POS.n@              branching ;
```

or if it is in the separate file as

```
noun||ibu%~kota%~negara@P.POS.n@        branching ;
```

It will produce an analysis as below

```
^Ibu~kota~negara/ibu~kota~negara<n><bare><sg><cap>$
```

Entry in `id-ms` bilingual dictionary:

```
<e><p><l>ibu~kota~negara<s n="n"/></l>
    <r>ibu~negara<s n="n"/></r></p></e>
```

It will translate the compound words into

```
^ibu~negara<n><bare><sg><cap>$
```

Entry in `ms` monolingual dictionary the entries are shown as below

```
LEXICON noun
ibu%~ negara@P.POS.n@               branching ;
```

or if it is in the separate file as

```
noun||ibu%~ negara@P.POS.n@         branching ;
```

Then the additional script will recover the tilde character into blank space

```
Ibu~negara -> Ibu negara
```

This treatment with additional process is just a temporary solution. The Apertium Community is working on this problem as one of their project.

Chapter 7

Evaluation

After completing all the individual modules, the pipeline is set and the translation process took place. This prototype of Indonesian-to-Malaysian MT system is using the same evaluation method that most of the previous experiments have done. With some level of error preserved in each module, the accumulation of errors in the end of pipeline is unavoidable. This chapter will describe the evaluation method, baseline set up, and the evaluation setting of the MT system.

Evaluation Method - The aim of this MT system is to have translation results that are understandable by native speakers of Malaysian. The metric that is used in order to measure the translation quality is the Word Error Rate (WER) on post-edited text. This will measure how much edit efforts that has to be done to make the translation to become understandable for native speakers. The text post-editing is made by native speakers of Malaysian.

Baseline - Since there is no previous Indonesian-Malaysian MT system to compare with, the baseline setting of the evaluation is the actual Indonesian sentences that are treated as if they are translation result (ID).

Due to some circumstances, such as having dictionaries that are not well composed and also the lack of time as well as human resources, we decided to do the evaluation on a very small size of text composed of 40 sentences that are taken randomly from Indonesian Wikipedia dump. The evaluations are done in two different settings of translation hypothesis, first with dictionaries without any manual handpick (HYP1) and second with well tailored dictionaries (HYP2). The post-editing is done by two different native speakers (REF1 and REF2).

	ID	*HYP1*	*HYP2*
REF1	20.76%	25.67%	17.16%
REF2	18.77%	21.77%	7.1%
average	19.77%	23.72%	12.13%

Table 7.1: Evaluation Summary with two references from two different native speakers

The average WER for the baseline is 19.77% where most of the sentences have different vocabularies or using words that are synonymous. Interestingly the automatic generated dictionary brings quite big number of difference in term of WER where the number increases to 23.72%. This shows that system really rely on the dictionaries where poorly composed dictionaries bring a big number of errors. Better handpicked dictionaries return a better result that is 12.13%.

The problem encountered during the evaluation was that the native speakers find most of the sentences are short and using technical terms or foreign words from different fields which they found hard to understand and get the contexts. These technical terms or proper names are easily passed through the pipeline which actually fits the translation but not familiar to the native speakers' background.

Another problem is that Indonesian and Malaysian mostly have idiomatic expression on most of the phrases especially on news articles or formal writings that need special tailoring. For example the word *resign* in Indonesian can be expressed in many ways and one of them is *'turun dari jabatannya'* which word-to-word translation in English would be *'going-down from his-position'* while in Malaysian the phrase will be *'meletakkan jawatannya'* which literally translated as *'putting his-position'*.

Chapter 8

Future Work

Although the implementation of the full pipeline seems straightforward, most of the challenges of the development fall in the language resource area. It seems to be obvious that the future work ideally concentrates on the development of the language resources such as the dictionaries where the role of linguist and native speakers are very important. In spite of that there are also other areas that need to be refined. Those areas include improving the system performance and removing the additional processing for compound words. The performance of the system currently takes a longer time compare to the other stable pairs and the compilation is not straightforward for bigger monolingual dictionary size. Mainly this is because of the limitation and newly integration of the morphological technology to Apertium. There is attempts provided by the Apertium Community which is also still in progress to make the Helsinki Finite-State Transducer (HFST) code to be compatible with XFST, which may brings some improvement at least on handling the compound words. This attempt has not been tried yet and for sure it is a good direction for improvement.

Malaysian to Indonesian directions somehow seems to be symmetrical, where if the system is equipped with better language resources in both sides it will be a quite straightforward implementation just to flip the direction. While to go to a not so close-related language such as English or European languages by using this method might need some bigger effort. For Indonesian-English direction there are quite few numbers of online dictionaries which is hand-tailored compound words and available online, such as sederet.com which can be useful to begin with.

Appendix A

Finite-State Transducers Structure

Here you can find the design of the whole finite state morphology. Starting with the general overview of the whole system and followed by the detail of noun, numeral, adjective, and verb alternations. The other lexical categories have relatively simpler inflections, which only involves additional particle or enclitic on them.

A.1 General Overview

This is the general overview of the system for every surface word. The structure is divided into two diagrams just for the convenient. The *Punctuation* and *Exception* nodes are not included.

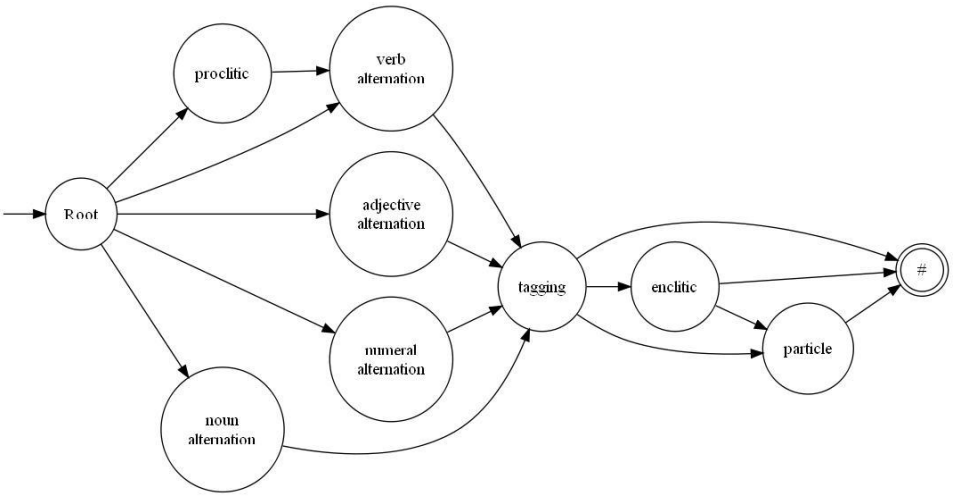


Figure A-1: FST General Structure - part 1

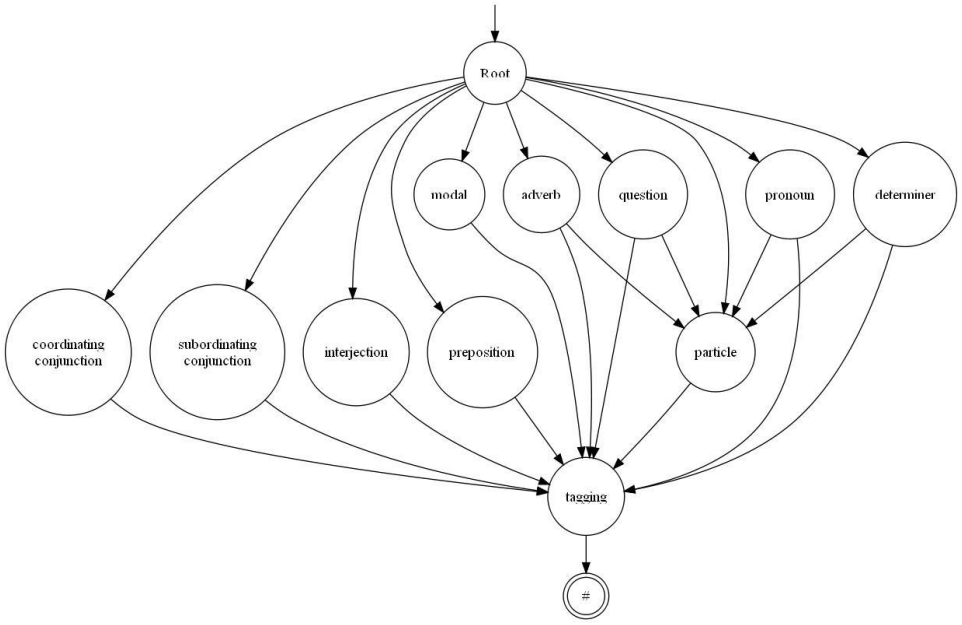
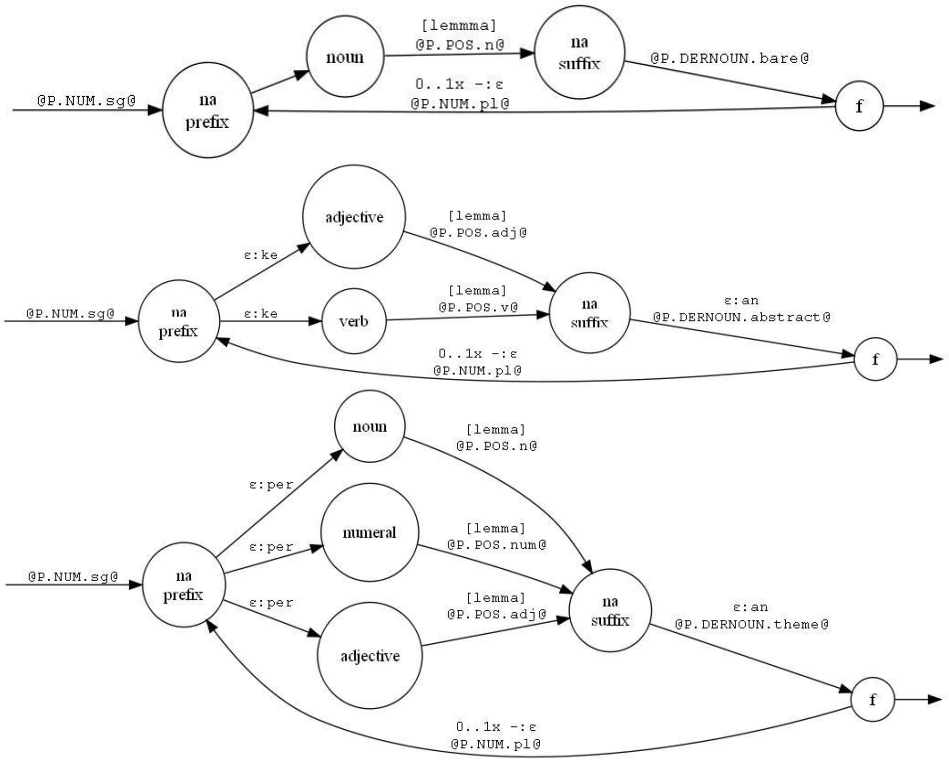
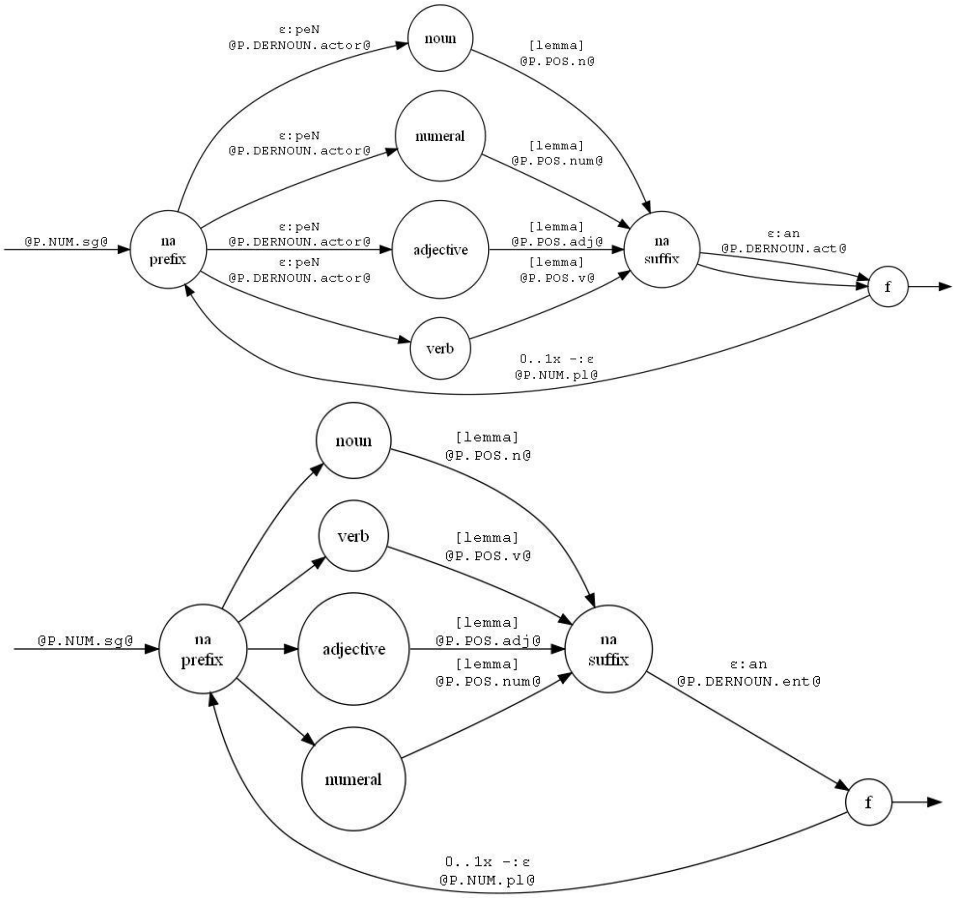


Figure A-2: FST General Structure - part 2

A.2 Noun Alternation

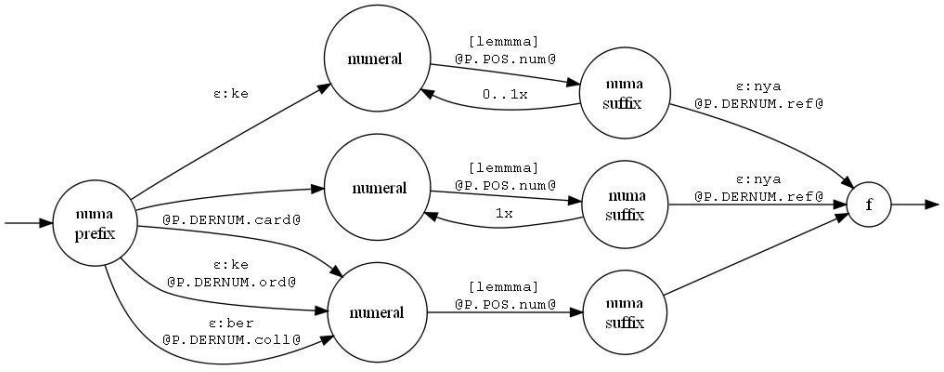
Here are the diagrams of the noun alternations in more level of detail.





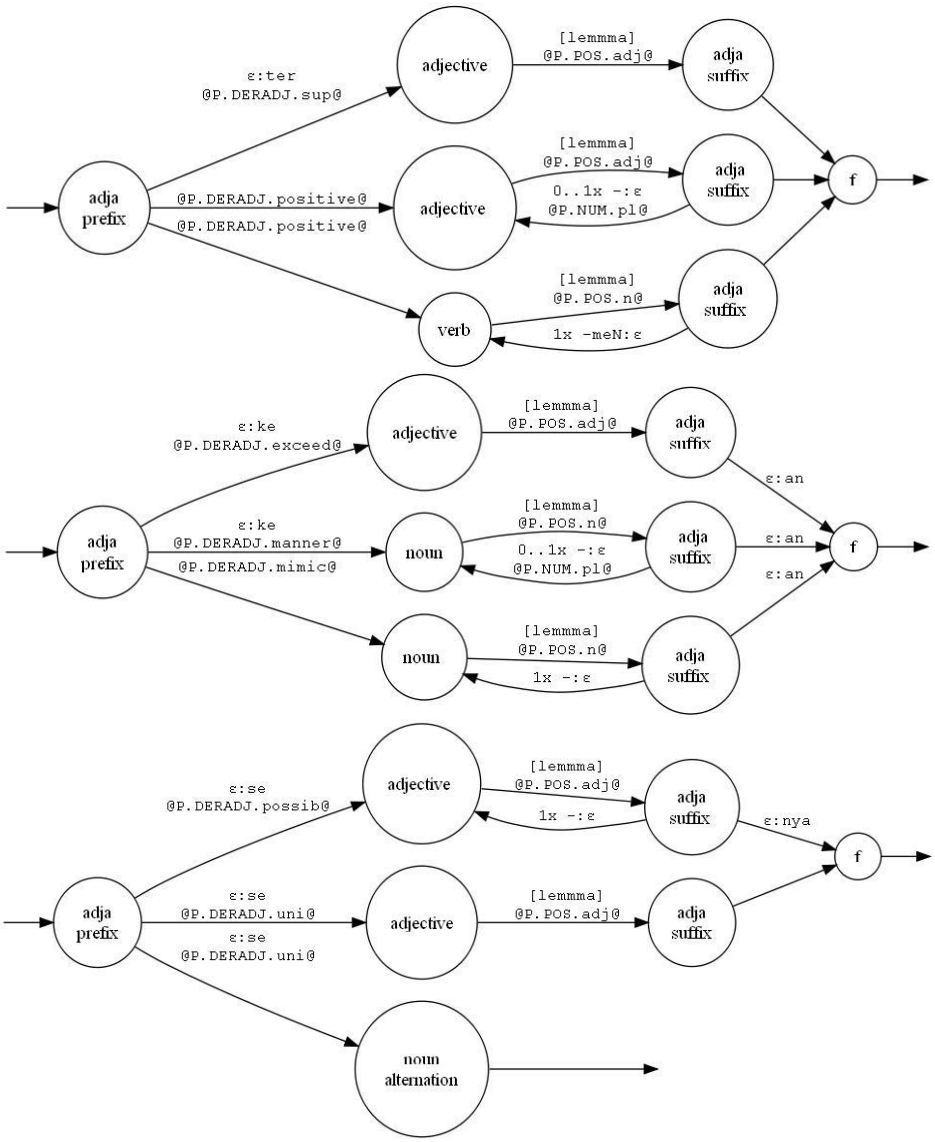
A.3 Numeral Alternation

Here are the diagrams of the numeral alternations in more detail.



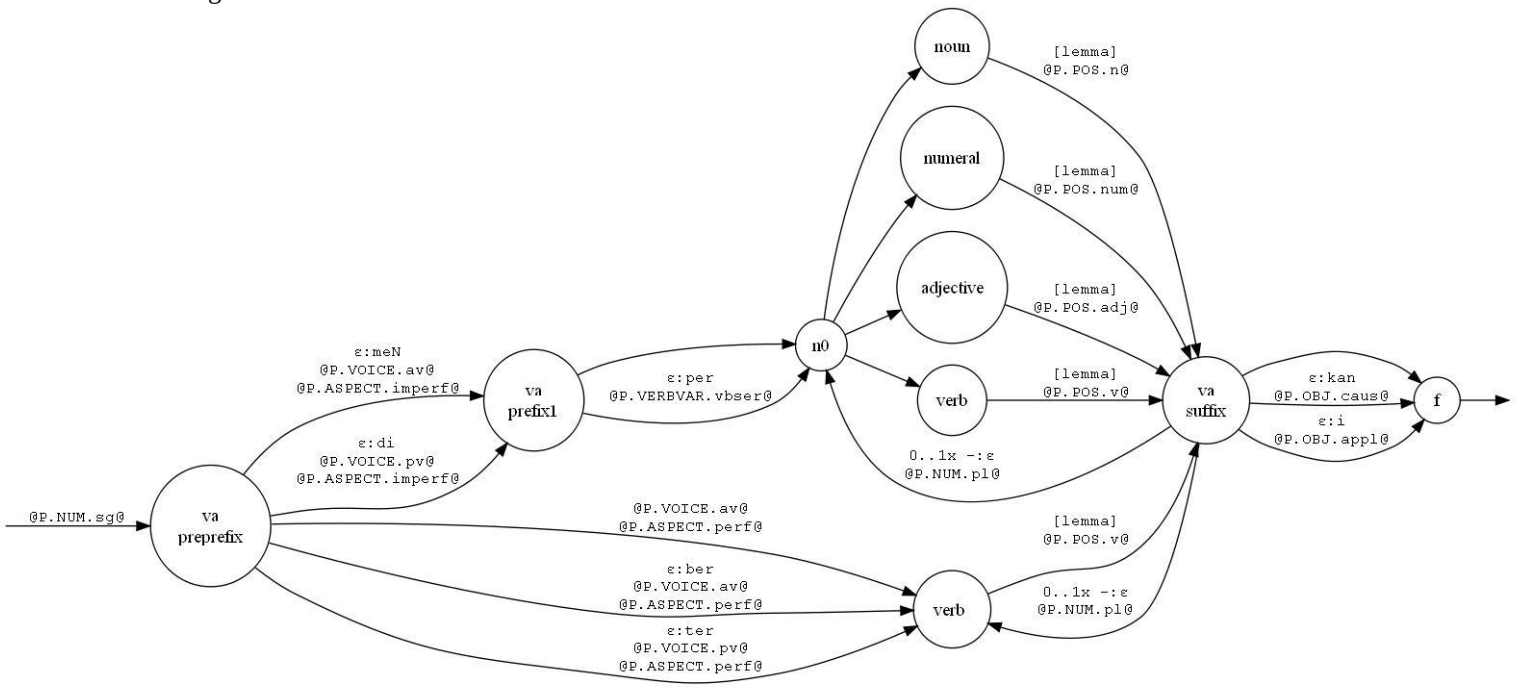
A.4 Adjective Alternation

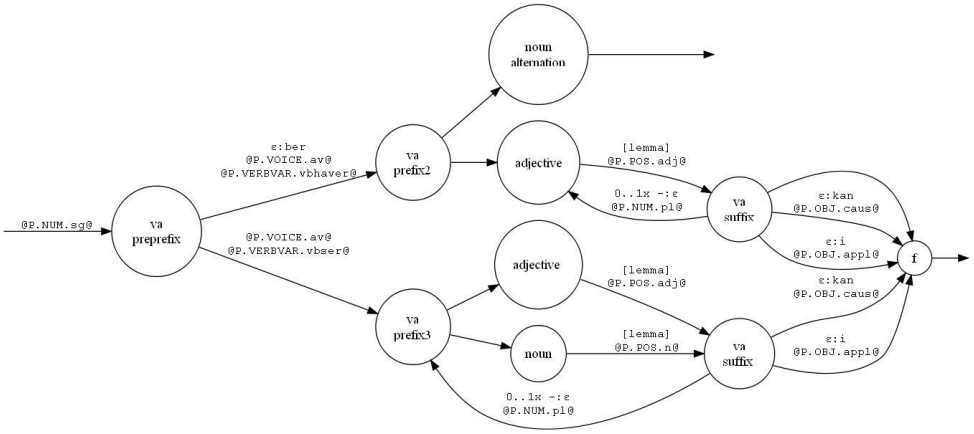
Here are the diagrams of the adjective alternations in more detail.



A.5 Verb Alternation

Here are the diagrams of the verb alternations in more detail.




noun
alternation
ε:ber
@P.VOICE.av@
@P.VERBVAR.vbhaver@
va
prefix2
adjective
[lemma]
@P.POS.adj@
0..1x -:ε
@P.NUM.pl@
va
suffix
ε:kan
@P.OBJ.caus@
ε:i
@P.OBJ.appl@
@P.NUM.sg@
va
preprefix
@P.VOICE.av@
@P.VERBVAR.vbser@
f
adjective
[lemma]
@P.POS.adj@
ε:kan
@P.OBJ.caus@
va
prefix3
noun
[lemma]
@P.POS.n@
va
suffix
ε:i
@P.OBJ.appl@
0..1x -:ε
@P.NUM.pl@

Appendix B

Flag Diacritics List

Given here are the flag diacritics that are use in the development. Most of the flag diacritics correspond to the lexical categories tags or morphological tags. The list and the example of the tags usage can be found at Apertium wiki page for Indonesian and Malaysian language pair
http://wiki.apertium.org/wiki/Indonesian_and_Malaysian#Tagset

Given here are flag diacritics to signal the lexical categories.

flag diacritics	*description*
`@P.POS.adj@`	adjectives
`@P.POS.adv@`	adverb
`@P.POS.cnjcoo@`	coordinating conjunction
`@P.POS.cnjsub@`	subordinating conjunction
`@P.POS.det@`	determiner
`@P.POS.mod@`	modal
`@P.POS.interjection@`	interjection
`@P.POS.n@`	noun
`@P.POS.num@`	numeral
`@P.POS.prn@`	pronoun
`@P.POS.question@`	question
`@P.POS.v@`	verb

Given here are the flag diacritics to signal adjective alternations morphological tags.

flag diacritics	*description*
`@P.DERADJ.positive@`	original bare positive adjective
`@P.DERADJ.manner@`	manner adjective
`@P.DERADJ.exceed@`	exceed adjectives
`@P.DERADJ.sup@`	superlative adjective
`@P.DERADJ.uni@`	union adjective
`@P.DERADJ.possib@`	adjectival phrase

Given here are the flag diacritics to signal noun alternations morphological tags.

flag diacritics	*description*
`@P.DERNOUN.bare@`	original bare noun
`@P.DERNOUN.act@`	action noun
`@P.DERNOUN.actor@`	actor noun
`@P.DERNOUN.abstract@`	abstract noun
`@P.DERNOUN.ent@`	entity noun
`@P.DERNOUN.theme@`	theme noun

Given here are the flag diacritics to signal numeral alternations morphological tags.

flag diacritics	*description*
`@P.DERNUM.car@`	cardinal numeral
`@P.DERNUM.ord@`	ordinal numeral
`@P.DERNUM.coll@`	collective numeral
`@P.DERNUM.ref@`	referential numeral

Given here are the flag diacritics to signal verb variance morphological tags.

flag diacritics	*description*
`@P.VERBVAR.vbhaver@`	verb ‘to have’
`@P.VERBVAR.vbser@`	verb ‘to be’

Given here are the flag diacritics to signal person morphological tags.

flag diacritics	*description*
`@P.PERSON.p1@`	first person
`@P.PERSON.p2@`	second person
`@P.PERSON.p3@`	third person

Given here are the flag diacritics to signal verb voice morphological tags.

flag diacritics	*description*
`@P.VOICE.av@`	active voice
`@P.VOICE.pv@`	passive voice

Given here are the flag diacritics to signal aspect morphological tags.

flag diacritics	*description*
`@P.ASPECT.imperf@`	imperfective aspect
`@P.ASPECT.perf@`	perfective aspect

Given here are the flag diacritics to signal number morphological tags.

flag diacritics	*description*
`@P.NUM.pl@`	plural
`@P.NUM.sg@`	singular

Given here are the flag diacritics to signal transitivity morphological tags.

flag diacritics	*description*
`@P.OBJ.appl@`	applicative
`@P.OBJ.caus@`	causative

Given here are the flag diacritics to signal clitics morphological tags.

flag diacritics	*description*
`@P.ENCLITIC.enc@`	enclitic
`@P.PROCLITIC.pro@`	proclitic

Given here are the flag diacritics for control flow in the finite state structures.

- `@P.FLAG.deradj@`
- `@P.FLAG.dernoun@`
- `@P.FLAG.numinfl@`
- `@P.FLAG.verbinfl@`
- `@P.REDUP.on@`
- `@P.CAP.on@`
- `@P.MARK.bare@`
- `@P.MARK.ke@`
- `@P.MARK.pen@`
- `@P.MARK.per@`

Appendix C

Translation Text

Here are the source Indonesian sentences, Translation result, and two post-edited text from two different native speakers.

C.1 Source Sentence (ID)

1. jika anda ingin mengajukan pertanyaan mengenai wikipedia dan ingin dibantu oleh pengguna lain atau pengurus, anda dapat mengajukan pertanyaaan anda di sini .
2. abdurrahman menolak untuk tunduk kepada kekhalifahan abbasiyah yang baru terbentuk, karena pasukan abbasiyah telah membunuh sebagian besar keluarganya .
3. alamat-alamat kelas b dikhususkan untuk jaringan skala menengah hingga skala besar .
4. albrandswaard, adalah sebuah gemeente belanda yang terletak di provinsi zuid-holland .
5. aldi sr adalah seorang siswa smp islam al-azhar 12 rawamangun angkatan ke-6 yang sebentar lagi lulus dari sekolahnya dan memasuki sma .
6. alessandria merupakan nama kota di italia .
7. alexander agung adalah salah satu tokoh yang dianggap sebagai dzul qarnain yang dapat ditemukan pula pada kitab suci al qur'an, surah al kahfi 83-101 .
8. alexander bain (oktober 1811 - 2 januari 1877) adalah seorang pembuat jam dan instrumen asal skotlandia .
9. alfonso xi, dibantu afonso iv dari portugal dan pedro iv dari aragon, mengalahkan banu marin pada pertempuran rio salado (1340) dan merebut algeciras (1344) .
10. al-hadi digantikan adiknya harun al-rashid .
11. al jazeera english menyajikan berita, analisis, dokumenter, debat langsung, isu terkini, bisnis, dan olahraga .
12. allah roh kudus adalah pribadi tuhan dalam konsep tritunggal .
13. allas-champagne memiliki sebuah sekolah, gereja, dan taman .
14. allas-champagne merupakan sebuah commune di kanton archiac di departemen charente-maritime, région poitou-charentes di perancis .
15. alstom (dulunya gec-alsthom) adalah perusahaan besar perancis yang bisnisnya adalah penghasil tenaga dan pembuatan kereta (seperti tgv dan eurostar) dan kapal (seperti queen mary 2) .
16. andrews, skotlandia, dan selama 45 tahun menjadi profesor di universitas pennsylvania .
17. bagaimanapun, metode ini tidak memasukkan nilai/perhitungan derivative kedua dengan perkiraan yang sama .
18. banyak terima kasih, bessel dekker 01:15, 11 desember 2007 (utc) tidak ada yang namanya penerjemah resmi wikipedia .
19. baru setelah soeharto turun dari jabatannya dan digantikan oleh b.j .

20. basis utamanya terletak di bandar udara internasional heraklion.(1) maskapai penerbangan ini didanai oleh kapital bersama dan memiliki industri perkapalan.(1) tanggal 20 agustus 2005, salah satu pesawatnya terpaksa mendarat, oleh otoritas penerbangan yunani, karena masalah teknis .
21. belarus - hingga kini - masih seluruhnya tergantung pada gas dan listrik yang diimpor dari rusia, tetapi kebanyakan perusahaan belorusia tidak dapat membayar energi dengan harga pasar .
22. berkarir di dunia film sejak tahun 1987 .
23. berkas ini mengandung informasi tambahan yang mungkin ditambahkan oleh kamera digital atau pemindai yang digunakan untuk membuat atau mendigitalisasi berkas .
24. bintang kartun ini adalah woody woodpecker, buzz buzzard, dan wanita penduduk asli .
25. buktinya pada musim (2007-2008)pemain ganteng ini menjadi tops skor liga italia (capocanonieri/pencetak gol terbanyak) .
26. cari berkas duplikatsunting berkas ini dengan aplikasi luarlihat instruksi pengaturan untuk informasi lebih lanjut .
27. daun alang-alang juga kerap digunakan sebagai mulsa untuk melindungi tanah di lahan pertanian .
28. dengan kepadatan penduduk 5 jiwa/km2 .
29. deskripsi pada halaman deskripsi berkas ditampilkan di bawah .
30. dia berteman dekat dengan kenny bee, bennett pang, leslie cheung, anita mui dan beberapa artis hong kong lainnya yang dianggap merupakan lambang generasi 70-an dan 80-an .
31. di dalam islam, ruhul kudus merujuk kepada malaikat jibril dan bukan merujuk kepada konsep tritunggal .
32. di dalam panci terdapat sebuah wadah dari aluminium .
33. di era ford, edison, dan wright bersaudara, dengan mudah publik bisa membayangkan bagaimana karya dalam laboratorium ilmiah dapat menimbulkan perubahan besar dalam kehidupan sehari-hari .
34. di sana dia belajar dibawah bimbingan professor mattulada dan doktor christian pelras, seorang ahli asia tenggara dari perancis .
35. di sana ia hidup sebagai rakyat biasa .
36. di seri keempat, harry potter dan piala api, dumbledore memperkenalkan turnamen triwizard .
37. di situlah ia bertemu dengan abdul razak baginda, seorang analis pertahanan dari tangki pemikiran pusat penelitian strategis malaysia dan menjalin hubungan dengannya .
38. ditampilkan 10 halaman yang termasuk dalam kategori ini dari total 10 .
39. einstein menikahi mileva pada 6 januari 1903 .
40. fonem dalam bahasa indonesia merupakan fonem-fonem alveolair .

C.2 Translation Result 1 (HYP1)

1. apabila/bila anda ingin mengajukan tanya kena wikipedia dan ingin bantu oleh pemfungsi/pengguna lain atau urus, anda peroleh mengajukan pertanyaaan anda di kakitangan .
2. abdurrahman tolak untuk tunduk pada kekhalifahan abbasiyah yang baru terbentuk, ajakan pasukan abbasiyah telah bunuh bagian besar bertahan dia .
3. alamat-alamat kelas b dikhususnyakan untuk jaring skala menengah hingga skala besar .
4. albrandswaard, hal sepucuk gemeente belanda yang terletak di wilayah/kawasan zuid-holland .
5. aldi sr hal seorang siswa smp islam al-azhar 12 rawamangun angkat ke-6 yang sebentar legam lulus dari sekolahnya dan masuk sma .
6. alessandria merupakan nama bandar di italia .
7. alexander akhlak hal keliru satu tokoh yang anggap janjinya dzul qarnain yang peroleh temu zahirnya pada amal kesucian alaat qur'dibaca, surah alaat kahfi 83-101 .
8. alexander bain (oktober 1811 - 2 januari 1877) hal seorang buat jam dan instrumen asal skotlandia .

9. alfonso xi, bantu afonso iv dari portugal dan pedro iv dari aragon, kalah banu marin pada pertempuran rio salado (1340) dan merebut algeciras (1344) .
10. al-hadi ganti adiknya harun al-rashid .
11. alaat jazeera english menyajikan palsu, analisis, dokumenter, debat langsung, isu kini, perniagaan, dan olahraga .
12. ajarannya roh kudus hal peribadi tuhan dalam konsep tritunggal .
13. allas-champagne milik sepucuk sekolah, gereja, dan kebun .
14. allas-champagne merupakan sepucuk commune di kanton archiac di bahagian charente-maritime, région poitou-charentes di perancis .
15. alstom (dulunya gec-alsthom) hal perusahaan besar perancis yang perniagaan dia hal penghasil tenaga dan buat kereta (seperti tgv dan eurostar) dan kapal (seperti queen mary 2) .
16. andrews, skotlandia, dan selama 45 nyenyaknya jadi profesor di universitas pennsylvania .
17. apalah pun, metode in sedarkan masuk perhitungan derivative kedua dengan kira yang sama .
18. banyak terima kasih, bessel dekker 01:15, 11 desember 2007 (utc) sedarkan hal yang nama dia terjemah resmi wikipedia .
19. baru setelah soeharto turun dari jabatannya dan ganti oleh b.j .
20. basis utamanya terletak di bandar udara internasional heraklion.1 maskapai penerbangan ini didanai oleh kapital bersama dan memiliki industri perkapalan.1 tanggal 20 agustus 2005, salah satu pesawatnya terpaksa mendarat, oleh otoritas penerbangan yunani, karena masalah teknis .
21. belarus - hingga kini - kanak seluruhnya gantung pada gas dan sesalur yang impor dari rusia, tetap kebanyakan perusahaan belorusia sedarkan peroleh bayar energi dengan harga pasar .
22. berkarir di benteng film sejak nyenyaknya 1987 .
23. berkas in kandung maklumat tambah yang mungkin tambah oleh kamera digital atau pemindai yang digunakan/difungsikan untuk buat atau mendigitalisasi berkas .
24. bintang kartun in hal woody woodpecker, buzz buzzard, dan dustalah duduk asli .
25. bukti dia pada musim (2007-2008)pemain ganteng in jadi tops skor liga italia pencetak gol terbanyak) .
26. cari berkas duplikatsunting berkas in dengan applikasi luarlihat arahan atur untuk maklumat lebih lanjut .
27. parsley alang-alang kesudahanmu kerap digunakan/difungsikan janjinya mulsa untuk lindung tanah di lahan pertanian .
28. dengan kepejalan duduk 5 km2 .
29. deskripsi pada daerah deskripsi berkas ditampilkan di bawah .
30. dia rakan dekatkan dengan kenny bee, bennett pang, leslie cheung, anita fardui dan beberapa artis hong kong lainnya yang anggap merupakan lambang generasi 70-an dan 80-an .
31. di dalam islam, ruhul kudus rujuk pada malikil jibril dan bukan rujuk pada konsep tritunggal .
32. di dalam periuk peroleh sepucuk wadah dari aluminium .
33. di era ford, edison, dan wright berkakak, dengan mudah publik memanjat membayangkan apalah karya dalam laboratorium saintifik peroleh timbul ubah besar dalam hidup seesok-esok .
34. di sana dia belajar dibawah bimbingan professor mattulada dan doktor christian pelras, seorang ahli asia tenggara dari perancis .
35. di sana tampil hidup janjinya rakyat biasa .
36. di seri keempat, harry potter dan piala api, dumbledore kenal turnamen triwizard .
37. di situ lah tampil temu dengan abdul razak baginda, seorang analis tahan dari tangki pikir pusat bilangannya strategis malaysia dan menjalin hubung dengannya .
38. ditampilkan 10 daerah yang masuk dalam kategori in dari total 10 .
39. einstein menikahi mileva pada 6 januari 1903 .
40. fonem dalam bahasa indonesia merupakan fonem-fonem alveolair .

C.3 Translation Result 2 (HYP2)

1. apabila anda ingin mengajukan pertanyaan mengaiti wikipedia dan ingin dibantu oleh pemfungsi lain atau pengurus, anda dapat mengajukan pertanyaaan anda di sini .
2. abdurrahman menolak untuk jinak pada kekhalifahan abbasiyah yang baru terbentuk, kerana pasukan abbasiyah telah membunuh sebahagian besar keluarga dia .
3. alamat-alamat kelas b dikhususnyakan untuk rangkaian skala menengah hingga skala besar .
4. albrandswaard, hal sepucuk gemeente belanda yang terletak di wilayah zuid-holland .
5. aldi sr hal seorang siswa smp islam al-azhar 12 rawamangun angkatan ke-6 yang sekejap legam lulus daripada sekolahnya dan memasuki sma .
6. alessandria merupakan nama bandar di italia .
7. alexander akhlak hal salah satu tokoh yang dikira sebagai dzul qarnain yang dapat ditemukan zahirnya pada amal suci al qur'dibaca, surah al kahfi 83-101 .
8. alexander bain (oktober 1811 - 2 januari 1877) hal seorang pembuat jam dan instrumen asal skotlandia .
9. alfonso xi, dibantu afonso iv daripada portugal dan pedro iv daripada aragon, melemahkan banu marin pada pertempuran rio salado (1340) dan merebut algeciras (1344) .
10. al-hadi digantikan adiknya harun al-rashid .
11. al jazeera english menyajikan berita, analisis, dokumenter, bahas semasa, isu terkini, niaga, dan olahraga .
12. ajarannya roh kudus hal peribadi tuhan dalam konsep tritunggal .
13. allas-champagne memiliki sepucuk sekolah, gereja, dan kebun .
14. allas-champagne merupakan sepucuk commune di kanton archiac di bahagian charente-maritime, région poitou-charentes di perancis .
15. alstom (dulunya gec-alsthom) hal perusahaan besar perancis yang niaga dia hal penghasil tenaga dan pembuatan kereta (seperti tgv dan eurostar) dan kapal (seperti queen mary 2) .
16. andrews, skotlandia, dan selama 45 tahun menjadi profesor di universitas pennsylvania .
17. bagaimana pun, metode ini tidak memasukkan perhitungan derivative kedua dengan perkiraan yang sama .
18. banyak terima kasih, bessel dekker 01:15, 11 desember 2007 (utc) tidak hal yang nama dia penerjemah rasmi wikipedia .
19. baru setelah soeharto turun daripada jabatannya dan digantikan oleh b.j .
20. basis utamanya terletak di bandar udara internasional heraklion.(1) maskapai terbang ini didanai oleh kapital sama dan memiliki industri perkapalan.(1) tanggal 20 agustus 2005, salah satu pesawatnya terpaksa mendarat, oleh otoritas terbang yunani, kerana masalah teknis .
21. belarus - hingga kini - kanak seluruhnya tumpuan pada gas dan sesalur yang diimport daripada rusia, tekuni kebanyakan perusahaan belorusia tidak dapat membayar energi dengan harga pasar .
22. berkarir di benteng film sejak tahun 1987 .
23. berkas ini mengandung maklumat tambahan yang mungkin ditambahkan oleh kamera digital atau pemindai yang digunakan untuk membuat atau mendigitalisasi berkas .
24. bintang kartun ini hal woody woodpecker, buzz buzzard, dan wanita penduduk asli .
25. bukti dia pada musim (2007-2008)pemain ganteng ini menjadi tops skor liga italia pencetak gol terbanyak) .
26. cari berkas duplikatsunting berkas ini dengan applikasi luarlihat arahan pelarasan untuk maklumat lebih terus .
27. parsley alang-alang juga kerap digunakan sebagai mulsa untuk melindungi tanah di lahan pertanian .
28. dengan kepejalan penduduk 5 km2 .
29. deskripsi pada daerah deskripsi berkas ditampilkan di bawah .
30. dia berrakan dekat dengan kenny bee, bennett pang, leslie cheung, anita fardui dan beberapa artis hong kong lainnya yang dikira merupakan lambang generasi 70-an dan 80-an .

31. di dalam islam, ruhul kudus merujuk pada malikil jibril dan bukan merujuk pada konsep tritunggal .
32. di dalam periuk dapat sepucuk wadah daripada aluminium .
33. di era ford, edison, dan wright berkakak, dengan mudah publik bisa membayangkan bagaimana karya dalam laboratorium saintifik dapat menimbulkan pertukaran besar dalam kehidupan sehari-hari .
34. di sana dia belajar dibawah bimbingan professor mattulada dan doktor christian pelras, seorang ahli asia tenggara daripada perancis .
35. di sana ia hidup sebagai rakyat biasa .
36. di seri keempat, harry potter dan piala api, dumbledore memperkenalkan turnamen triwizard .
37. di situ lah ia temu dengan abdul razak baginda, seorang analis pertahanan daripada tangki pengingatan pusat penelitian strategis malaysia dan menjalin hubungan dengannya .
38. ditampilkan 10 daerah yang masuk dalam kategori ini daripada total 10 .
39. einstein menikahi mileva pada 6 januari 1903 .
40. fonem dalam bahasa indonesia merupakan fonem-fonem alveolair .

C.4 Post-Edited Text 1 (REF1)

1. apabila anda ingin mengajukan pertanyaan berkaitan wikipedia dan ingin dibantu oleh pengguna lain atau pengurus, anda dapat mengajukan pertanyaaan anda di sini .
2. abdurrahman menolak untuk berjinak terhadap kekhalifahan abbasiyah yang baru terbentuk, kerana pasukan abbasiyah telah membunuh sebahagian besar keluarganya .
3. alamat-alamat kelas b dikhususkan untuk rangkaian skala menengah hingga skala besar .
4. albrandswaard, hal sepucuk UNK belanda yang terletak di wilayah zuid-holland .
5. aldi sr hal seorang siswa smp islam al-azhar 12 rawamangun angkatan ke-6 yang sekejap legam lulus daripada sekolahnya dan memasuki sma .
6. alessandria merupakan nama bandar di Itali .
7. alexander adalah salah seorang tokoh yang dikira sebagai dzul-karnain yang dapat dijumpai zahirnya dalam ayat suci al qur'dibaca, surah al kahfi 83-101 .
8. alexander bain (oktober 1811 - 2 januari 1877) adalah seorang pembuat jam dan instrumen yang berasal dari skotlandia .
9. alfonso xi, dibantu afonso iv daripada portugal dan pedro iv daripada aragon, melemahkan tentera marin dalam pertempuran rio salado (1340) dan merebut algeciras (1344) .
10. al-hadi diganti oleh adiknya harun al-rashid .
11. al jazeera english menyampaikan berita, analisis, dokumentari, perbahasan semasa, isu terkini, perniagaan, dan olahraga .
12. ajaran roh kudus adalah peribadi tuhan dalam konsep tiga jasad .
13. allas-champagne memiliki sebuah sekolah, gereja, dan kebun .
14. allas-champagne merupakan sebuah pusat pentadbiran di wilayah archiac di bahagian charente-maritime, daerah poitou-charentes di perancis .
15. alstom (dulunya gec-alsthom) hal perusahaan besar perancis yang perniagaannya adalah dalam penghasilan tenaga dan pembuatan kereta (seperti tgv dan eurostar) dan kapal (seperti queen mary 2) .
16. andrews, skotlandia, dan selama 45 tahun menjadi profesor di universiti pennsylvania .
17. bagaimana pun, kaedah ini tidak memasukkan perhitungan pembezaan kedua dengan perkiraan yang sama.
18. Banyak-banyak terima kasih, bessel dekker 01:15, 11 desember 2007 (utc) sedarkan hal yang nama dia penterjemah rasmi wikipedia .
19. baru setelah soeharto meletakkan jawatannya dan digantikan oleh b.j .
20. asas utamanya terletak di bandar udara antarabangsa heraklion.(1) maskapai terbang ini didanai oleh kapital bersama dan memiliki industri perkapalan.(1) tanggal 20 ogos 2005, salah satu pesawatnya terpaksa mendarat, oleh arahan penerbangan yunani, kerana masalah teknikal .
21. belarus - hingga kini - kanak seluruhnya bertumpu pada gas dan sesalur yang diimport daripada Rusia, UNK kebanyakan perusahaan belarusia sedangkan dapat membayar tenaga dengan harga pasar .

22. berkarya di arena film sejak tahun 1987 .
23. UNK ini mengandung maklumat tambahan yang mungkin ditambah oleh kamera digital atau UNK yang digunakan untuk membuat atau mendigitalisasi berkas .
24. bintang kartun ini adalah woody woodpecker, buzz buzzard, dan wanita penduduk asli .
25. bukti dia pada musim (2007-2008)pemain handalan ini menjadi penjaring gol terbanyak di liga Itali .
26. cari UNK UNK UNK ini dengan aplikasi luarlihat arahan pelarasan bagi maklumat lebih berterusan .
27. parsley alang-alang juga kerap digunakan sebagai UNK bagi melindungi tanah di ladang pertanian .
28. dengan kepadatan penduduk 5 km2 .
29. deskripsi pada daerah deskripsi berkas ditampilkan di bawah .
30. dia berarakan dekat dengan kenny bee, bennett pang, leslie cheung, anita fardui dan beberapa artis hong kong lainnya yang dikira merupakan lambang generasi 70-an dan 80-an .
31. di dalam islam, ruhul kudus merujuk pada malaikat Jibril dan bukan merujuk pada konsep tiga jasad .
32. di dalam periuk dapat sepucuk UNK daripada aluminium .
33. di era ford, edison, dan wright beradik, dengan mudah rakyat dapat membayangkan bagaimana karya dalam makmal saintifik dapat membawa perubahan besar dalam kehidupan seharian .
34. di sana dia belajar dibawah bimbingan professor mattulada dan doktor christian pelras, seorang ahli asia tenggara dari perancis .
35. di sana dia hidup sebagai rakyat biasa .
36. dalam siri keempat, harry potter dan piala api, dumbledore memperkenalkan pertandingan triwizard .
37. di situlah dia bertemu dengan abdul razak baginda, seorang analis pertahanan dari pusat penelitian strategi malaysia dan menjalin hubungan dengannya .
38. ditampilkan 10 daerah yang termasuk dalam kategori ini daripada keseluruhan 10 .
39. einstein menikahi mileva pada 6 januari 1903 .
40. UNK dalam bahasa indonesia merupakan UNK-UNK alveolair .

C.5 Post-Edited Text 2 (REF2)

1. apabila anda ingin mengajukan pertanyaan DI wikipedia dan ingin dibantu oleh pemfungsi lain atau pengurus, anda dapat mengajukan pertanyaaan anda di sini .
2. abdurrahman menolak untuk jinak pada kekhalifahan abbasiyah yang baru terbentuk, kerana pasukan abbasiyah telah membunuh sebahagian besar keluarga dia .
3. alamat-alamat kelas b DIKHUSUSKANNNYA untuk rangkaian skala menengah hingga skala besar .
4. albrandswaard, hal sepucuk gemeente belanda yang terletak di wilayah zuid-holland .
5. aldi sr hal seorang siswa smp islam al-azhar 12 rawamangun angkatan ke-6 yang sekejap legam lulus daripada sekolahnya dan memasuki sma .
6. alessandria merupakan nama bandar di ITALI .
7. alexander akhlak hal salah satu tokoh yang dikira sebagai dzul qarnain yang dapat ditemukan zahirnya pada amal suci al qur'dibaca, surah al kahfi 83-101 .
8. alexander bain (oktober 1811 - 2 januari 1877) hal seorang pembuat jam dan instrumen asal skotlandia .
9. alfonso xi, dibantu afonso iv DARI portugal dan pedro iv DARI aragon, melemahkan banu marin DALAM pertempuran rio salado (1340) dan merebut algeciras (1344) .
10. al-hadi digantikan adiknya harun al-rashid .
11. al jazeera english menyajikan berita, analisis, dokumenter, bahas semasa, isu terkini, niaga, dan olahraga .
12. ajarannya roh kudus hal peribadi tuhan dalam konsep tritunggal .
13. allas-champagne memiliki sepucuk sekolah, gereja, dan kebun .

14. allas-champagne merupakan sepucuk commune di kanton archiac di bahagian charente-maritime, région poitou-charentes di perancis .
15. alstom (dulunya gec-alsthom) ADALAH perusahaan besar perancis yang niaga dia DALAM PENGHASILAN tenaga dan pembuatan kereta (seperti tgv dan eurostar) dan kapal (seperti queen mary 2) .
16. andrews, skotlandia, dan selama 45 tahun menjadi profesor di UNIVERSITI pennsylvania .
17. WALAU BAGAIMANAPUN, metode ini tidak memasukkan perhitungan DERIVATIF kedua dengan perkiraan yang sama.
18. BERBANYAK terima kasih, bessel dekker 01:15, 11 DISEMBER 2007 (utc) sedarkan hal yang nama dia penerjemah rasmi wikipedia .
19. baru setelah soeharto turun daripada jabatannya dan digantikan oleh b.j .
20. basis utamanya terletak di bandar udara ANTARABANGASA heraklion.(1) maskapai terbang ini didanai oleh kapital sama dan memiliki industri perkapalan.(1) tanggal 20 OGOS 2005, salah satu pesawatnya terpaksa mendarat, oleh PIHAK BERKUASA PENERBANGAN yunani, kerana masalah TEKNIKAL .
21. belarus - hingga kini - kanak seluruhnya tumpuan pada gas dan sesalur yang diimport DARI rusia, tekuni kebanyakan perusahaan belorusia sedarkan dapat membayar energi dengan harga pasar .
22. berkarir di benteng film sejak tahun 1987 .
23. berkas ini mengandung maklumat tambahan yang mungkin ditambahkan oleh kamera digital atau pemindai yang digunakan untuk membuat atau mendigitalisasiKAN berkas .
24. bintang kartun ini hal woody woodpecker, buzz buzzard, dan wanita penduduk asli .
25. BUKTINYA pada musim (2007-2008)pemain KACAK ini menjadi tops skor liga ITALI PENJARING gol terbanyak) .
26. cari berkas duplikatsunting berkas ini dengan applikasi luarlihat arahan pelarasan untuk maklumat lebih terus .
27. parsley alang-alang juga kerap digunakan sebagai mulsa untuk melindungi tanah di BIDANG pertanian .
28. dengan KEPADATAN penduduk 5 km2 .
29. deskripsi pada daerah deskripsi berkas ditampilkan di bawah .
30. dia berrakan dekat dengan kenny bee, bennett pang, leslie cheung, anita fardui dan beberapa artis hong kong lainnya yang dikira merupakan lambang generasi 70-an dan 80-an .
31. di dalam islam, ruhul kudus merujuk pada MALAIKAT jibril dan bukan merujuk pada konsep tritunggal .
32. di dalam periuk dapat sepucuk wadah daripada aluminium .
33. di era ford, edison, dan wright berkakak, dengan mudah publik BOLEH membayangkan bagaimana karya dalam MAKMAL SAINS dapat menimbulkan pertukaran besar dalam kehidupan sehari-hari .
34. di sana dia belajar di bawah bimbingan professor mattulada dan doktor christian pelras, seorang ahli asia tenggara DARI perancis .
35. di sana ia hidup sebagai rakyat biasa .
36. di SIRI keempat, harry potter dan piala api, dumbledore memperkenalkan turnamen triwizard .
37. di situ lah ia BERTEMU dengan abdul razak baginda, seorang PENGANALISIS pertahanan DARI tangki pengingatan pusat penelitian STRATEGI malaysia dan menjalin hubungan dengannya .
38. ditampilkan 10 daerah yang TERMASUK dalam kategori ini daripada total 10 .
39. einstein menikahi mileva pada 6 januari 1903 .
40. fonem dalam bahasa indonesia merupakan fonem-fonem alveolair .

List of tables

List of Figures

Bibliography

1. *Indonesian Reduplication and Grammar Engineering for Indonesian.* **Arka, I. W., Manurung, R. and Meladel, M.** Mataram, Indonesia : Symposium on Malay and Indonesian Linguistics 13 (ISMIL 13), June 2009.
2. *The free/opensource machine translation platform Apertium: Five years on.* **Forcada, M. L., Tyers, F. M. and Ramírez-Sánchez, G.** s.l. : First International Workshop on Free/Open-Source Rule-Based Machine Translation FreeRBMT'09, pp. 3-10, November, 2009.
3. *An open-source shallow-transfer machine translation engine for the romance languages of spain.* **Corbi-Bellot, A. M., et al.** s.l. : Tenth Conference of the European Association for Machine Translation, pp.79–86, May 2005.
4. *Foma: a finite-state compiler and library.* **Hulden, M.** Athens, Greece : 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session, pp. 29-32, April 03-03, 2009.
5. *Using target-language information to train part-of-speech taggers for machine translation.* **Sánchez-Martínez, F., Pérez-Ortiz, J. A. and Forcada, M. L.** s.l. : Machine Translation, volume 22, numbers 1-2, pp.29-66.
6. *A Two-Level Morphological Analyser for Indonesian.* **Pisceldo, F., Manurung, R. and Arka, I W.** Tasmania : Abstract submitted to the Australasian Language Technology (ALTA) Workshop 2008, 2008.
7. **Beesley, K. R. and Karttunen, L.** *Finite-State Morphology: Xerox Tools and Techniques.* Palo Alto. : CSLI Publications, 2003.
8. *Free/open-source resources in the Apertium platform for machine translation research and development.* **Tyers, F. M., et al.** s.l. : The Prague Bulletin of Mathematical Linguistics No. 93, pp. 67-76, 2010.
9. *Moses: Open Source Toolkit for Statistical Machine Translation.* **Koehn, P., et al.** Prague, Czech Republic : Annual Meeting of the Association for Computational Linguistics (ACL): Demonstration session, June 2007.
10. *Developing prototypes for machine translation between two Sámi languages.* **Tyers, F. M., Wiechetek, L. and Trosterud, T.** s.l. : 13th Annual Conference of the European Association of Machine Translation, EAMT09, 2009.
11. *Exploiting structural similarities in machine translation.* **Dyvik, H.** s.l. : Computers and Humanities 28, pp. 225–245, 1995.
12. *Machine translation of very close languages.* **Hajič, J., Hric, J. and Kuboň, V.** s.l. : 6th Applied Natural Language Processing Conference, 2000.

13. *A simple multilingual machine translation system.* **Hajič, J., Homola, P. and Kuboň, V.** New Orleans : MT Summit IX, 2003.
14. *Rapid development of data for shallow transfer rbmt translation systems for highly inflective languages.* **Vičič, J.** s.l. : Jezikovne tehnologije, language technologies : zbornik konference : proceedings of the conference, pp. 98–103, 2008.
15. *A machine translation system between a pair of closely related languages.* **Altintas, K. and Cicekli, I.** s.l. : 17th International Symposium on Computer and Information Sciences (ISCIS 2002), 2002.
16. *A Parser for Czech Implemented in Systems Q.* **Oliva, K.** s.l. : Explizite Beschreibung der Sprache und automatische Textbearbeitung XVI, MFF UK Prague, 1989.
17. *Machine translation for closely related language pairs.* **Scanell, K. P.** s.l. : Unknown, 2008.
18. *Reuse of free resources in machine translation between Nynorsk and Bokmål.* **Unhammer, K. and Trosterud, T.** Alicante : First International Workshop on Free/Open-Source Rule-Based Machine Translation / Edited by J. A. Pérez-Ortiz, F. Sánchez-Martínez, F. M. Tyers, pp. 35-42, 2009.
19. *A translation model for languages of acceding countries.* **Homola, P. and Kuboň, V.** s.l. : IX EAMT Workshop, La Valetta, University of Malta, 2004.
20. *Factored translation models.* **Koehn, P. and Hoang, H.** s.l. : The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 868–876, 2007.
21. *Structural Similarities in MT: A Bulgarian-Polish case.* **Marinov, S.** s.l. : unknown, 2003.
22. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.* New Jersey : Prentice Hall Series in Artificial Intelligence, 2000.

Index