Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus

Septina Dian Larasati, Vladislav Kuboň, and Daniel Zeman

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{larasati,vk,zeman}@ufal.mff.cuni.cz

Abstract. This paper describes a robust finite state morphology tool for Indonesian (MorphInd), which handles both morphological analysis and lemmatization for a given surface word form so that it is suitable for further language processing. MorphInd has wider coverage on handling Indonesian derivational and inflectional morphology compared to an existing Indonesian morphological analyzer [1], along with a more detailed tagset. MorphInd outputs the analysis in the form of segmented morphemes along with the morphological tags. The implementation was done using finite state technology by adopting the two-level morphology approach implemented in Foma. It achieved 84.6% of coverage on a preliminary stage Indonesian corpus where it mostly fails to capture the proper nouns and foreign words as expected initially.

1 Introduction

Indonesian, or Bahasa Indonesia as the locals would call it, is the official language of Indonesia. The language is spoken by approximately 230 million people throughout the country with only 23 million native speakers. Typologically, the language could be partially classified as isolating and partially as agglutinative.

Language technology research on this language has been quite enthusiastic in recent years but without having a well developed continuous long term plan. There are many language tools such as a parser, a semantic analyzer and a speech recognition tool. Our Indonesian Morphology tool, MorphInd, was intended to set up proper ground work before doing any further language processing. MorphInd is applied to enrich a raw Indonesian text with morphological information, a preprocessing stage of developing an Indonesian corpus.

MorphInd was inspired by an existing Indonesian morphological analyzer tool [1] (hereinafter called IndMA), where we found that the analysis produced was inadequate. More on this matter described further in section 2. MorphInd introduces a more fine-grained tagset compared to IndMA and gives the output in form of segmented morphemes as an added value. In addition to that, the lemmata are also tagged independently for lemmatization purposes.

The goal of the work described in this paper is to have an Indonesian morphology tool which has broader morphological and lexical coverage, provided with richer and less ambiguous linguistic information in the analysis, and tested on common Indonesian

text. The work includes: the design of the new tagset to cope with Indonesian morphological phenomena; the format of the analysis output which includes constructing the morphemic segmentation format and lemma marking; and a better organization of the lexical categories. The coverage of the tool is then evaluated on an Indonesian corpus which consists of text coming from different domains.

2 Motivation

The work in Indonesian Morphology was done over a long period. There was previous work on developing an Indonesian stemmer [2,3]. The limitation of these tools is that they only recover the root of an affixed surface form without any additional linguistic information, which could be encoded by the occurence or combination of morphemes. Then an initial version of a morphological analyzer [4] developed in PC-KIMMO was introduced. Unfortunately, reduplication, which is one of Indonesian morphology's crucial points was not yet covered by the tool.

The latest work on the morphological analyzer is a finite state tool [1] implemented on XFST [5], a commercial finite state technology (FST) toolkit. The tool is able to handle most of Indonesian morphosyntactic and morphophonemic phenomena. In spite of how robust it models the morpheme's composition, the linguistic information that it produces in the analysis was rather simple and ambiguous. Although it was developed in a FST environment, the reduplication and affixed reduplication are also covered by the tool (more about this matter in section 3.4)

We decided IndMA was a good starting point to develop MorphInd, which is basically a refinement of IndMA, although there were some major changes on the finite state architecture that was taken when we ported the rules. Those changes are intended to make it more organized for further development. Here we point out four issues that we found as the limitations of IndMA, which we refined in MorphInd.

Issue #1. Shallow Lexical Categorizations. It was designed with only a simple tagset that consists of four major different lexical tags namely 'Noun', 'Verb', 'Adjective', and 'Etc' plus several additional language feature tags. This shallow categorization is not adequate to be passed onto another tool such as a parser, where categories such as 'Numeral', 'Adverb' and many others play an important role.

Issue #2. Underspecified Analysis. The output is in the form of a lemma followed by morphological tags. There are some problems of underspecified analysis since the output is in a simple form with a limited tagset. There is the same analysis for different word derivations. Figure 1 shows examples of the verb v.*kirim* (send/deliver) derivation, where several derived words have the same lemma and the derived words falls in the same lexical category.

On the other hand, the generation step of the analysis into the surface forms, outputs many varieties of morpheme combinations allowed by the finite state network, where many of them are invalid surface word forms (see figure 2).

Input		Output
v. *kirim*	(v. send/deliver)	`kirim+Verb`
n. *kiriman*	(n. packages)	`kirim+Noun`
n. *pengirim*	(n. deliverer)	`kirim+Noun`
n. *pengiriman*	(n. delivery)	`kirim+Noun`

Fig. 1. IndMA analysis examples

Input	Output	
`kirim+Noun`	n. *kiriman*	(n. packages)
`kirim+Noun`	n. *pengirim*	(n. deliverer)
`kirim+Noun`	n. *pengiriman*	(n. delivery)
`kirim+Noun`	**pemberkiriman*	
`kirim+Noun`	**perkiriman*	
`kirim+Noun`	**kerberkiriman*	
`kirim+Noun`	**kekiriman*	

** invalid surface forms*

Fig. 2. IndMA generation examples

Issue #3. Morphosyntactic Rules. The morphosyntactic rules that were defined in IndMA cover almost all possible cases in Indonesian, disregarding the exceptionals, which are not trivial to solve. But there are more morphosyntactic cases which are trivial to solve, that are not covered by IndMA, such as clitics.

Issue #4. Software license. IndMA was developed on XFST, which is a commercial finite-state automata and transducer. The tool uses a patent encumbered function which does the non-concatenative morphology operation for the reduplication, therefore the overall software cannot be used freely. The aim for MorphInd is to make it available for any individuals who want to utilize or refine the tool.

3 Tool Design

MorphInd was designed to address the four issues that were previously mentioned. MorphInd produces analysis that only covers morphology phenomena; it does not handle syntax, but its output can be used as input to many other Natural Language Processing (NLP) tasks. MorphInd analyzes tokens as unigrams and does not take into account any neighbouring tokens. MorphInd does not return any syntactical functions on the analyses, although some functions are easily recognized by the word order or the clitics. For example, we do not mark the 'subject' of the sentence where it can be easily recognized by a common proclitic that is attached to a verb, but the fact that the surface word form has a pronoun proclitic is kept in the analyses. We decided to do this since this processing will be the task of a parser. In a more complex system, MorphInd can be used as one of the modules that gives morphological tags before parsing.

3.1 Tagset Design and Lexical Category Organization

MorphInd organizes the lexical entries into 17 different lexical categories. Those categories are basically 'Noun', 'Verb', 'Adjective' as in IndMA and we broke down 'Etc' into several lexical categories such as 'Preposition' and 'Modal', where most of these categories are closed word classes with entries that are easy to enumerate manually. These categories correspond to a lemma lexical category tag, which tag the lemma for lemmatization purposes.

MorphInd also has a fine-grained tagset which was inspired by the PENN Treebank tagset and adapted it accordingly to Indonesian morphology. The tagset also adopts the concept of positional tags of the Prague Dependency Treebank tagset to cope with most of the language behaviours that occur simultaneously in a surface word. The tagset contains morphological tags in three positions and a lemma tag. The first position reflects the actual lexical category of the surface word, while the second and third tag positions are there to give more specific linguistic information. Table 1 gives the complete tagset.

3.2 Analysis Format

We decided to make the output in the form of segmented morphemes, which it shows how the morphemes combined. This will make the output more precise and less ambiguous for the generation step. The surface word form was segmented to its morphemes. The lemma is directly followed by a lemma tag, which corresponds to the first position of the word form tag, and that they are distinguished by lowercase. Lemma tag can differ from the first position of the same token, because of derivation (see figure 3). This format will make it easier to extract the lemma if needed (e.g., `kirim<vb>`). Then the sequence of the whole segmented morphemes including the lemma tag are followed by morphological tags as described in the tagset. Clitics, as they stand as independent words semantically, are treated as a single word form which has its own analysis but they are glued in the surface word's overall analysis as one of the morphemes. In this way, the fact that the morpheme was clitic is still kept in the output. Figure 3 shows several word derivation output examples and a word phrase of the lemma v. *kirim* (v. send/deliver) with clitics.

Input		**Output**
v. *kirim*	(v. send/deliver)	`kirim<vb><VB><SG><AV>`
v. *mengirim*	(v. send/deliver)	`meN+kirim<vb><VB><SG><AV>`
n. *kiriman*	(n. package)	`kirim<vb>+an<NN><SG>`
n. *pengiriman*	(n. delivery)	`peN+kirim<vb>+an<NN><SG>`
ph. *kumengirimkannya*	(ph. I send/deliver him/her)	`aku<prp><PRP><SG><1>+meN +kirim<vb>+kan<VB><SG><AV> +dia<prp><PRP><SG><3>`

Fig. 3. MorphInd Derivation Analysis Examples

Table 1. MorphInd Tagset

1st position		2nd position		3rd position	
NN	Noun	PL	Plural	F	Feminine
NNP	Proper noun	SG	Singular	M	Masculine
				D	Non-Specified
PRP	Personal pronoun	PL	Plural	1	First Person
		SG	Singular	2	Second Person
				3	Third Person
VB	Verb	PL	Plural	AV	Active Voice
		SG	Singular	PV	Passive Voice
CD	Numeral	C	Cardinal Numeral		
		O	Ordinal Numeral		
		D	Collective Numeral		
CC	Coordinating conjunction				
SC	Subordinative conjunction				
JJ	Adjective	P	Positive		
		S	Superlative		
FW	Foreign word				
IN	Preposition				
MD	Modal				
DT	Determiner				
RB	Adverb				
RP	Particle				
NEG	Negation				
UH	Interjection				
COP	Copula				
WH	Question				
UNK	Unknown				

3.3 Morphosyntactic and Morphophonemic Operation

Indonesian is not an inflected language as Slavic languages are, although there are several morphemes that bring language features such as verb conjugation to mark active and passive voices or noun declination to mark the gender (this inflection is not produced anymore and doesm not relate to the grammar such as word gender agreement). Indonesian is a mildly agglutinative language when compared to Finnish or Turkish where the morpheme-per-word ratio is higher. There are several common subject or object pronouns of the sentence event that can be represented as clitics (proclitic and enclitic). Most of the morphological phenomena are word derivational cases done by concatenative affixation operations (prefix, suffix, circumfix, and infix). These affixation examples can be seen in figure 3.

Input		**Output**
n. *gerigi*	(n. teeth)	`gerigi<nn><NN><PL>`
n. *gigi-gigi*	(n. teeth)	`gigi<nn><NN><PL>`
n. *2 buku*	(n. 2 books) (lit n. *2 book)	`2<cd><CD><C> buku<nn><NN><SG>`
n. *dua buku*	(n. two books) (lit n. *two book)	`dua<cd><CD><C> buku<nn><NN><SG>`
n. *buku-buku*	(n. books)	`buku<nn><NN><PL>`
n. **2 buku-buku*	(lit n. two books)	`2<cd><CD><C> buku<nn><NN><PL>`

Fig. 4. MorphInd plural form examples

Input		**Output**
num. *2*	(num. 2)	`2<cd><CD><C>`
num. *dua*	(num. two)	`dua<cd><CD><C>`
num. *ke-2*	(num. second)	`ke+2<cd><CD><O>`
num. *kedua*	(num. second)	`ke+dua<cd><CD><O>`

Fig. 5. MorphInd numeral alternation examples

MorphInd handles infixations differently by putting the surface word as one of the entries in the dictionary, since infixations are not common anymore in Indonesian. For example the word n. *gerigi* (n. teeth), which it is the word n. *gigi* (n. tooth) with *er* infix in *g+er+igi* arrangement, are defined in the dictionary and marked as plural. The word n. *gerigi* is not common anymore and has the word n. *gigi-gigi* as its equivalent word. Both analyses of the examples can be seen at figure 4. There are no feature agreements except numerical agreement for a noun to have singular form if preceeded by a plural numeral e.g., *dua buku* (*lit.* *two book). In this case MorphInd only works in the level of single word tokens and does not capture the plurality of the whole phrase. Given in figure 5 are also examples of numeral alternations.

Deriving Nouns, Adjectives, Verbs, and Numerals are the most productive derivational morphosyntactic and morhophonemic operations. It also includes the non-concatenative morphology operation i.e. reduplication that occurs to mark the plural mood. We designed the finite state architecture into a more organized way, seperating the alternation based on those categories and on their affixation segments. The schema (without reduplication) is provided in table 2.

Table 2. Nouns, adjectives, verbs, and numerals alternation schema

		Preprefix	**Prefix**	**Lemma**	**Suffix**
	ε+		ε+		+ε
Noun	anti+		peN+		+an
Alternation	antar+		ke+	[lemma]	+wan
			per+		+wati
			ke+tidak+		
	ε+		ε+		+ε
Adjective	non+		ter+		+an
Alternation			ke+	[lemma]	+nya
			se+		
		ε+	ε+		+ε
Verb		meN+	per+		+kan
Alternation		di+		[lemma]	+i
		ber+			
Numeral			ε+		+ε
Alternation			ke+	[lemma]	+nya
			ber+		+belas

We reuse the morphophonemic rules from IndMA since those rules cover most of the cases. We ported and organized all the morphosyntactic rules. In addition to that, we added more rules, such as more affix concatenative rules, handling the clitics (proclitics and enclitics), additional particles (e.g., *-lah, -kah, -tah,* and *-pun*), and several additional compound word morphemes (e.g., *antar-* and *anti-*). The general MorphInd finite state schema can be found in figure 6.

3.4 Software License

IndMA uses the `compile-replace()` function provided by XFST to handle reduplication. It copies the marked morpheme that is going to be reduplicated during the compilation. This function is patent-encumbered which limits the usage of the tool. To loosen the license we decided not to use that function but tweaked the reduplication process. We also use Foma toolkit [6] instead of XFST to compile the tool so that MorphInd is suitable to fall into an Open Source license.

Foma, which falls under GNU General Public License, works in the similar way as XFST and accept XFST/LEXC code therefore several parts of the source code of

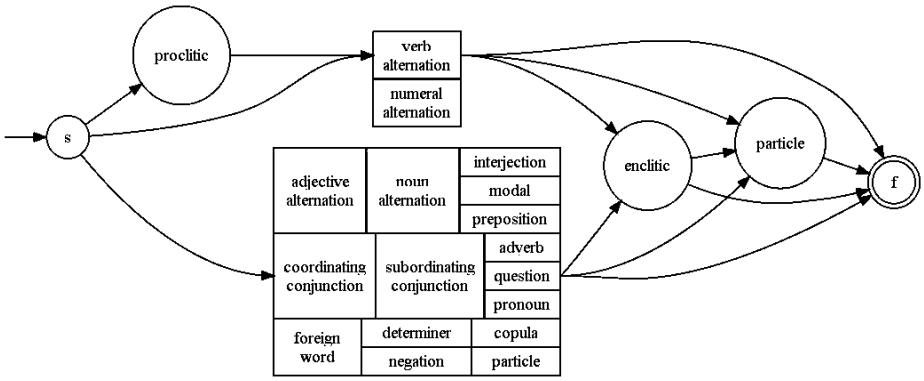

Fig. 6. MorphInd general finite state architecture schema

IndMA can be easily reused as needed. The tweaking is done by pairing all the marked morphemes with anything and discards all the pairs which are not similar. This causes the finite state network compilation time and memory consumption to explode. To handle this we limit the lexical entry size by splitting it into several parts and compile it as separate networks. All the resulting finite state networks then are wrapped together by a Perl script to build the tool.

4 Research on Indonesian

4.1 Linguistic Tools

The work on developing language resources for Indonesian is not enthusiastic compared to the work on developing lingustics tools. There were works on developing an Indonesian online dictionary [7] but its resources are not freely available. The entries are equipped with linguistic and anthropological information. There is also a project on developing an Indonesian wordnet [8] that is still ongoing.

While on the other hand, development of Indonesian linguistics tools are surprisingly popular and done with different approaches. Beside works on a morphological analyzer, there are also works on developing an Indonesian probabilistic part-of-speech tagger [9,10]. On the syntactic level, there are works on developing an Indonesian rule-based parser using PC-PATR [11], which relies on annotated lexical entries. Later, this tool is also being used to model a probabilistic parser learned from the parsed trees that it produces [12]. Even though the ground work for further processing is not properly established yet, it does not stop the researchers from trying to make semantic tools. There are also some work on semantics such as semantic analyzers [13,14].

4.2 Indonesian Corpus Plan

Since there are no available Indonesian lingusitic corpora, we initiatively collected Indonesian texts and prepared them for further linguistic processing. Although it is not

required to be a parallel corpus, we prefer to have Indonesian text that is aligned with English text. Mainly we collected the texts from the PAN Localization project output [15] and subtitles. Currently the Indonesian part consists of 45,011 sentences. The statistic of the corpus sources are given in figure 7.

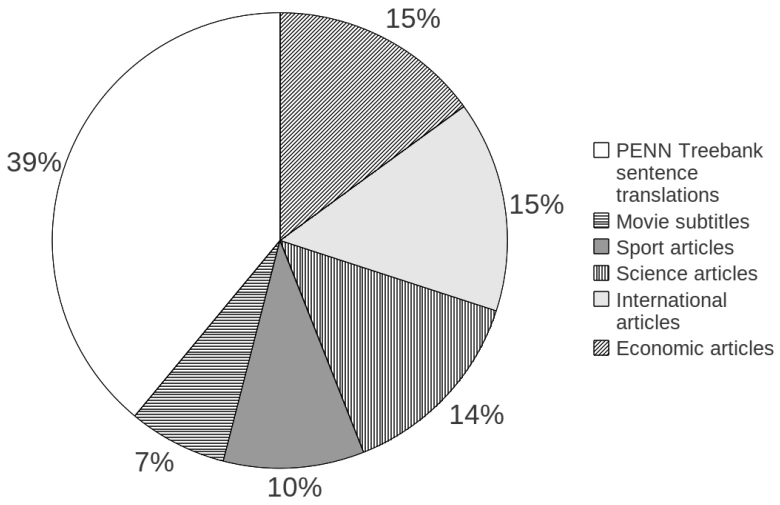

Fig. 7. Indonesian Parallel Corpus Source Statistic

For the initial plan, the final corpus will be in XML format following the PML schema [16] with several different layers such as morphology, syntax, etc. MorphInd will fill the morphology layer of the corpus. As the plan continues we are hoping to have an Indonesian-English parallel treebank corpus.

5 Evaluation

Test Set. We ran MorphInd and IndMA on Indonesian text that we have collected to measure the coverage. We made two types of test sets i.e. 5,000 sentences (T5K) and 10,000 sentences (T10K). There are nine sets of T5K and four sets of T10K. The sentences in a test set were chosen randomly without replacement from the text that we have collected (see section 4.2).

Metric. We used coverage as our metric. The coverage is measured in two ways, *overall* and *unique*. *Overall* is the ratio of the number of words analyzed and the number of the words in the text. *Unique* is the ratio of different word forms analyzed and the number of different word forms in the text.

Experiments. MorphInd consists of 3,954 lexical entries divided into 17 lexical categories. We did not port all the entries that are available in IndMA which has more entries but with several of them overlapped across the categories or in affixed forms. We also rebuilt IndMA to have the same lexical entries as MorphInd to make a comparable experiment (hereinafter called IndMA-Comparable). Detail of the tools' lexical entries can be found in table 3 and 4. The resulting comparison of the three tools can be seen in table 5.

Table 3. MorphInd lexical entries

Noun	2,222	Numeral	19	Particle	4
Verb	924	Adjective	576	Negation	3
Personal pronoun	13	Foreign word	0	Interjection	11
Coordinating Conjunction	3	Preposition	16	Copula	3
Subordinating Conjunction	32	Modal	6	Question	6
Determiner	15	Adverb	89	**TOTAL**	**3,942**

Table 4. IndMA and IndMA-comparable lexical entries

	IndMA	**IndMA-Comparable**
Noun	5,863	2,222
Verb	3,417	924
Adjective	19,036	576
Etc	4,153	220
TOTAL	**32,469**	**3,942**

MorphInd failed to outperform IndMA in *Unique* coverage since the number of lexical entries greatly differs and MorphInd lexical entries do not include proper nouns and foreign words. But with a good selection of the lexical entries, by chosing the most frequent and productive lemmas, MorphInd's *Overall* coverage became greater than IndMA. This is because MorphInd mainly covers clitics, numeral alternation, and additional particle morphemes which were not covered by IndMA. This can be easily seen on MorphInd's and IndMA-Comparable's results, where MorphInd had a better coverage with the same lexical entries.

6 Conclusion and Future Work

MorphInd produces robust morphological information in the output format i.e. morphemic segmentation, lemma morpheme position, lexical category, and morphological feature. The new robust tagset with broader categorization that it uses is also suitable

Table 5. Evaluation

	Test Sets	# Sentences	Overall	Unique
MorphInd	T5K	5,000	84.69±0.28	50.77±0.70
	T10K	10,000	84.61±0.10	47.19±0.35
IndMA	T5K	5,000	83.62±0.27	54.95±0.76
	T10K	10,000	83.46±0.06	51.39±0.05
IndMA-Comparable	T5K	5,000	81.91±0.18	44.60±0.66
	T10K	10,000	81.82±0.06	40.83±0.31

for a further language processing such as parsing. MorphInd gives a better coverage compared to IndMA.

The most current version of MorphInd can be found at the MorphInd homepage which includes MorphInd documentation, binaries, and source code.[1]

Yet for future improvements, we will investigate more morpheme behaviour to add to MorphInd, such as morphoponemic affixation exceptions on one syllable words. As its initial plan, this tool will enrich the morphological layer of the Indonesian corpus. We also will build an initial parser based on MorphInd's output.

7 Acknowledgement

This project was financially supported by the grant LC536 Centrum Komputační Lingvistiky of the Czech Ministry of Education.

References

1. Pisceldo F., Mahendra R., Manurung R., and Arka I W.: A Two-Level Morphological Analyser for Indonesian. Abstract submitted to the Australasian Language Technology (ALTA) Workshop 2008, Tasmania (2008)
2. Siregar, N.: Pencarian Kata Berimbuhan pada Kamus Besar Bahasa Indonesia dengan menggunakan Algoritma Stemming. Undergraduate thesis, Faculty of Computer Science, University of Indonesia (1995)
3. Adriani, M., Jelita A., Nazief, S B., Tahaghoghi M. and Williams H.: Stemming Indonesian: A Confix-Stripping Approach. ACM Transactions on Asian Language Information Processing, Vol. 6, No. 4 (2007)
4. Hartono, H.: Pengembangan Pengurai Morfologi untuk Bahasa Indonesia dengan Model Morfologi Dua Tingkat Berbasiskan PC-KIMMO. Undergraduate thesis, Faculty of Computer Science, University of Indonesia (2002)
5. Beesley, K.R., Karttunen, L.: Finite State Morphology. CSLI Publications, Palo Alto, CA (2003)
6. Hulden M.: Foma: a finite-state compiler and library. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session, pp. 29–32. Athens, Greece (2009)

[1] `http://ufal.ms.mff.cuni.cz/~larasati/MorphInd.html`

7. Pusat Bahasa: Kamus Besar Bahasa Indonesia Daring, http://pusatbahasa.diknas.go.id/kbbi/ (2008). Last Access: 14th February 2011.
8. Darma Putra, D. Arfan, A. Manurung R.: Building an Indonesian Wordnet. The Second International MALINDO Workshop (2008) http://bahasa.cs.ui.ac.id/wordnet/ Last Access: 14th February 2011.
9. Pisceldo F., Manurung, R., Adriani, M.: Probabilistic Part-of-Speech Tagging for Bahasa Indonesia. The Third International MALINDO Workshop, colocated event ACL-IJCNLP 2009, Singapore, August 1 (2009)
10. Farizki Wicaksono A., Purwarianti, A.: HMM Based Part-of-Speech Tagger for Bahasa Indonesia. The Fourth International MALINDO Workshop. Jakarta, Indonesia (2010)
11. Joice: Pengembangan lanjut pengurai struktur kalimat bahasa indonesia yang menggunakan constraint-based formalism. Undergraduate thesis, Faculty of Computer Science, University of Indonesia (2002)
12. Hari Gusmita, R., Manurung R: Some Initial Experiments with Indonesian Probabilistic Parsing. The Second International MALINDO Workshop (2008)
13. Dian Larasati, S., Manurung R.: Towards a Semantic Analysis of Bahasa Indonesia for Question Answering. In: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING) (2007)
14. Mahendra, R., Dian Larasati, S., Manurung R.: Extending an Indonesian Semantic Analysis-based Question Answering System with Linguistic and World Knowledge Axioms. In: Proceedings of the 22nd Pacific Asia Conference on Language, Information, and Computation (PACLIC 2008), pp. 262–271 (2008)
15. PAN Localization, http://www.panl10n.net/english/OutputsIndonesia2.htm. Last Access: 14th February 2011.
16. Prague Markup Language (PML), http://ufal.mff.cuni.cz/jazz/pml/index_en.html. Last Access: 14th February 2011.