Research Report on Local Language Computing: Development of Indonesian Language Resources and Translation System

Country Component	INDONESIA	**Report no.**	1
Phase no.	1	**Report Ref no.**	

Prepared By: Dr. Mirna Adriani
& Dr. Hammam Riza **Designation: _____**

Project Leader: Dr. Mirna Adriani **Signature: _____**

Date: Novermber 4, 2008

TABLE OF CONTENTS

Title

Local Language Computing: Development of Indonesian Language Resources and Translation System

Dr. Mirna Adriani, Faculty of Computer Science, University of Indonesia, Depok Campus, Depok 16424, Indonesia. Phone: (+62-21) 7863419. Fax: (+62-21) 7863415. Email: mirna@cs.ui.ac.id

Dr. Hammam Riza, IPTEKnet-BPPT.Jln. M.H. Thamrin 8, Gd.1 Lt.16, Jakarta 10340, Indonesia. Phone +62-21-316-8701. Email: hammam@iptek.net.id

Abstract

This first report contains our early study in developing language resources and translation system for Bahasa Indonesia. We report our literature survey on part-of-speech tagger and machine translation that has been developed on other languages. We also report our plan in building Indonesian corpus and translating PennTreebank corpus into Indonesian.

Introduction

Bahasa Indonesia is the national language of Indonesia which is spoken by more than 222 millions people (Etnologue, 2005). It is widely used in Indonesia to communicate in school, government offices, etc. Bahasa Indonesia became the formal language of the country, uniting its citizens who speak different languages. There are more than 742 distinct languages in the country (Etnologue, 2005) which are still used widely in various regions of the country. Bahasa Indonesia has become the language that bridges the language barrier among Indonesians who have different mother-tongues.

Within the context of international and regional economy, opportunities are also visible for Indonesia to accelerate its economic growth as the ASEAN and other countries in the Asia Pacific region have jointly agree to reduce trade barriers. The establishments of free trade areas such AFTA and APEC will stimulate further economic growth in the region through the increase of investment and trade traffic. Indonesia can capitalize on this opportunity by establishing positive business environment. However, the efforts to capitalize the opportunities to accelerate economic growth are hindered due to the lack ability of Indonesia in communicating with foreigner. MT rapidly being developed due to the increase of international communication and globalization, it is expected to bridge the language gap among nations in the world. It is said that the most significant development of the last decade in Language Technology is the appearance of commercial MT systems. In order to disseminate information available only in one language, automatic translation is required so that this information can be translated to many other languages quickly and accurately.

In this project we concentrate in developing a machine translation to translate English documents into Bahasa Indonesia. There are several components that are needed to be developed to prepare the research in machine translation, such as the parallel corpus and the part-of-speech tagger. Although the history of corpora is relatively short the technological advances enabled creating many different types of such sets of texts, so nowadays apart from monolingual corpora there are bilingual or multilingual corpora. Another type is called sample corpora and those show a state of language at a given point in time. A Reference corpus is one that can reliably portray all the features of a language. There are also historical corpora which

aim at comparison of past forms of a language with its present state; they can be subdivided into two kinds depending of the features they want to emphasize. Thus, diachronic corpora present samples of language with intervals of about a generation of users, while monitor corpora attempt to follow the language change while it occurs. Apart from historical approaches there are topic corpora which focus on a particular field of interest or a **genre**.

This is the first design criteria for the Indonesian reference corpus that we are building for PANL10n project. Genres of spoken and written texts are being intensively studied from various angles, e.g., communication studies, discourse analysis, computational linguistics, without arriving at a generally accepted definition (Aston and Burnard, 1998). Many corpora have been built to represent the language, but very few large corpora indicate genres, and when they do the typology of genres varies widely. For instance, the Brown corpus famously uses 15 textual categories, from press reportage (a text genre) to religion or skills and hobbies (domains), while the British National Corpus (BNC) uses 70 classes, such as academic or non-academic scientific texts or biography. Interestingly, genre classes in the BNC are an add-on proposed by David Lee (Lee, 2001) after the corpus construction, rather than a basic criterion of the corpus creation. The genre attribute was included in a few collections used in information retrieval (TREC HARD 2003 and 2004, or TREC-2006 Blog Track), but the set of genres proposed was either debatable (e.g. the 'reaction' genre in TREC HARD 2003), or limited to a single genre (e.g. the blog genre in TREC-2006 Blog Track).

The web is new, so it is even less not clear how to apply traditional notions of genre to web documents. In corpus-based genre studies, the main tendency has been to build one's own genre collection according to subjective criteria for corpus composition, genre annotation, and genre granularity. Genre annotation has been based either on the common sense of a single rater, or on the agreement of few annotators. In brief, as it is now, web genre analyses remain self-contained and corpus-dependent. Building a reference corpus of web genres is certainly difficult because web documents are often characterized by a high level of genre hybridism, by a fragmentation of textuality across several documents, by the impact of technical features such as hyperlinking, posting facilities and multi-authoring. Since the web is a huge reservoir of documents that can be easily mined for building all sorts of corpora, it is important to overcome the subjectivity that characterizes genre-related issues, in order to create sharable resources. What should we consider when designing a reference corpus of web genres? Genres of web documents show some traits that are not accounted for in TREC collections or in the BNC and that are, instead, important on the web.

According to Sharof [Sharoff, 2008], the BNC is a jungle when compared to smaller Brown-type corpora, but it looks more like an English garden when compared to the Internet. He compare English and Russian Internet corpora against their human-collected counterparts (BNC and RNC) using two methods: the first involves manual annotation of a subset of Internet corpora; the second one uses probabilistic classifiers. The study shows that the Internet is not radically different from the BNC: Internet corpora do contain a wide range of genres and approximate many genres that exist in their printed form, the same is true for the audience level (texts for professional or layman texts).

Besides the parallel corpus, we also need to develop Part-of-Speech Tagger (POSTAG) tools for Bahasa Indonesia. POSTAG labels each word according to its part-of-speech in the sentence such as noun, verb, etc. POSTAG is useful in eliminating sense ambiguity of a word in sentences.

POSTAG is usually built manually by linguists, however it is very expensive to tag words in a large collection. So it would be better if the tagging process can be done automatically. There have been many studies regarding POSTAG especially in English words, especially with the need to do the tagging faster and more accurate. For example, TAGGIT, is the first rule-based POS tagger [Greene & Rubin, 1971]. The POSTAG techniques have implemented in many languages such as POSTAG in Hindi, Telugu, and Bengali [Avinesh & Karthik, 2007] based on Transformation Based Learning, POSTAG for French [Chanod & Tapainen, 1995] and German [Brants, 2008] based on Hidden Markov Model etc.

POS tagger in a language is influenced by the characteristics of the language. As an example, a POS tagger for English classifies a verb into three categories according to its tenses

while in Bahasa Indonesia that kind of classification is not needed since Bahasa Indonesia does not distinguish tenses. So the POS tagger for one language is different from that for another language depending on the characteristics of the language. The tagset or the number of the tags that are used in the POSTAG is also different depending on the language's sentence structure. In English there are several tagsets that are available such as the Penn Treebank tagset [Marcus et.al., 1993] with 45 tags; Brown Corpus tagset [Francis & Kucera, 1982] with 87 tagset; Lancaster UCREL C5 with 61 *tag;*, and Lancaster C7 with 145 *tags*. Some of the tags that are used in Brown Corpus tagset may or may not be used in other tagsets, for example tag 'AT' which means article is used in Brown Corpus however it does not exist in Penn Treebank tagset.

There are many approaches that are known to build a POST tagger. In the beginning, most of POSTAGs are based on the rules of Greene & Rubin, 1971. However, this approach is not efficient and expensive for building the rules. Later, there has been another approach that is based on the statistical method, e.g., probability values, such as POSTAG that is developed based on Hidden Markov Model [Kupiec et.al., 1992], and Conditional Random Field [Lafferty et al., 2001; McCallum, 2003; Wallach, 2004; Avinesh & Karthik, 2007]. These statistical approaches produce good result if the tagger is trained using big collections without the use of any linguistic concepts. Brill tagger was developed based on the Transformation Based Learning [Brill, 1992] that considers not only probabilistic models but also linguistic concepts, where the approach learns the rules from the training texts.

The POSTAG for Bahasa Indonesia will use the appropriate tags based mainly on tags available in the Penn Treebank tagset, although it will also consider other tagsets, where appropriate, according to the characteristics of the language.

Methods

1.Design of Corpus

Based on Sharof study [Sharoff, 2008], we decided to classify the Indonesian corpus into 2 genres, which related to source of the corpus. We also adopt the notion of Domain for classifying the field (area) of interest which the corpus described, such as Business and Sport. Table 1 shows the genres and domains which we classify all the source documents of the corpus.

Table 1. Corpus Sources: Domain and Genre

		Economy	Sport	National	International	Science & Technolgy
Genre	News	ANTARA News Database	ANTARA News Sport	ANTARA News Database - National	ANTARA News Database - International	National Geography Indonesia Berita IPTEK
	Web	detikfinance.com mediaindonesia.com economy.okezone.com kompas.com/bisniskeuangan lampungpost.com tempo.co.id/bisnis liputan6.com/ekbis	detiksport.com mediaindonesia.com sports.okezone.com kompas.com/olahraga lampungpost.com tempo.co.id/olahraga liputan6.com/olahraga	detiknews.com mediaindonesia.com news.okezone.com kompas.com/nasional lampungpost.com tempo.co.id/nasional liputan6.com	detik.com mediaindonesia.com internasional.okezone.com kompas.com/internasional lampungpost.com tempo.co.id/internasional liputan6.com/luarnegri	detikinet.com mediaindonesia.com techno.okezone.com tekno.kompas.com lampungpost.com tempo.co.id/teknologi netsains.com

1.1. Creative Commons IPR

During the first quarter, our team have look into the legal issues of creating Indonesian corpus. Based on the information in Wikipedia (http://en.wikipedia.org/wiki/Creative_Commons_licenses), and also observe the way Wikipedia Commons (refer to http://commons.wikimedia.org/wiki/Main_Page) which shape up the licensing of Wikimedia

Projects. It is said that all text on all Wikimedia projects, except for Wikinews, is owned by the original writer and licensed under the GNU Free Documentation License (GFDL) and Wikinews is licensed under the Creative Commons Attribution 2.5.

We specifically study the policy for collecting the Indonesian corpus from various available sources, such as news agency, online media publisher, internet blog, websites and others. The original set of licenses grant the "baseline rights". The details of each of these licenses depend on the version and comprise a selection of four conditions:

Attribution (by): Licensees may copy, distribute, display and perform the work and make derivative works based on it only if they give the author or licensor the credits in the manner specified by these.

Noncommercial or **NonCommercial** (nc): Licensees may copy, distribute, display, and perform the work and make derivative works based on it only for noncommercial purposes.

No Derivative Works or **NoDerivs** (nd): Licensees may copy, distribute, display and perform only verbatim copies of the work, not derivative works based on it.

ShareAlike (sa): Licensees may distribute derivative works only under a license identical to the license that governs the original work. (See also copyleft.)

Mixing and matching these conditions produces sixteen possible combinations, of which eleven are valid "Creative Commons Licenses" (CC). Of the five invalid combinations, four include both the "nd" and "sa" clauses, which are mutually exclusive; and one includes none of the clauses. The five of the eleven valid licenses that lack the Attribution element have been phased out because 98% of licensors requested Attribution, but are still available for viewing on the website. There are thus six regularly used licenses:

1. Attribution alone (by)
2. Attribution + Noncommercial (by-nc)
3. Attribution + NoDerivs (by-nd)
4. Attribution + ShareAlike (by-sa)
5. Attribution + Noncommercial + NoDerivs (by-nc-nd)
6. Attribution + Noncommercial + ShareAlike (by-nc-sa)

Work licensed under a Creative Commons Licenses is protected by copyright applicable law. This allows CC licenses to be applied to all work protected by copyright law, including: books, plays, movies, music, articles, photographs, blogs, and websites. However, the license may not modify the rights allowed by fair use or fair dealing or exert restrictions which violate copyright exceptions. Furthermore, Creative Commons Licenses are non-exclusive non-revocable. Any work or copies of the work obtained under a Creative Commons license may continue to be used under that license.

In the case of works protected by multiple Creative Common Licenses, the user may choose either. In order to use the news database from **ANTARA News Agency**, we have submit a request letter for releasing the copyrighted newswire into CC license of type 5, **Attribution + Noncommercial + NoDerivs (by-nc-nd).**

For the articles which are collected from various **online media and websites**, we have submitted a request letter to each organization for transferring the copyright into CC license type 6, **Attribution + Noncommercial + ShareAlike (by-nc-sa).** The following table shows the

contact information of some online publisher. In order to use the documents from **E-book**, we have submit a request letter to the author for releasing the copyright into CC license of type 5, **Attribution + Noncommercial + NoDerivs (by-nc-nd)**

Table 2. Corpus Sources Contact Information

No	Source	Publisher
1	mediaindonesia.com	Kompleks Delta Kedoya, Jl. Pilar Raya Kav. A-D, Kedoya Selatan, Kebon Jeruk, Jakarta Barat - 11520 Telepon : (021) 5812088 (Hunting), Fax : (021) 5812102, 5812105 (Redaksi) E-mail : redaksi@mediaindonesia.co.id
2	kompas.com	PT. Kompas Cyber Media Gedung Kompas Gramedia, Unit II Lt. 5 Jl. Palmerah Selatan No. 22 – 28 Jakarta 10270, Indonesia. Telp : 62-21 5350377 / 5350388 Fax : 62-21 5360678 Redaksi: redaksikcm@kompas.co.id redaksikcm@kompas.com
3	okezone.com	Gedung Bimantara Lt. 4 Jln. Kebon Sirih Kav. 17-19 Jakarta 10340 Phone: 021 3902275 Fax: 021 3902295 Email: Redaksi@]okezone.com
4	detik.com	Aldevco Octagon Building Lantai 2 Jl. Warung Buncit Raya No.75 Jakarta Selatan 12740 Telp: (021) 794.1177 (Hunting) Fax: (021) 794.4472 redaksi@staff.detik.com
5	tempo.co.id	Redaksi Tempo Interaktif Kebayoran Center Blok A11 - A15 Jl. Kebayoran Baru - Mayestik, Jakarta 12440 Telp (021) 725 5625 Faks (021) 725 5645 interaktif@tempo.co.id
6	lampungpost.com	Jl. Soekarno Hatta 108 Rajabasa Bandar Lampung redaksilampost@yahoo.com

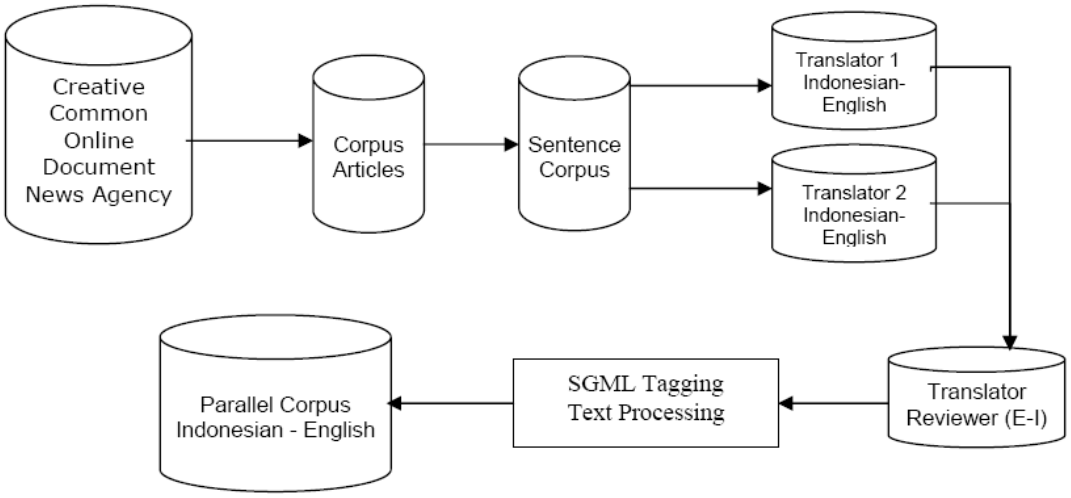
1.2. Corpus Collection Process

Activity of parallel text corpus Indonesian-English is a bilingual corpus collection from various sources as described in Section 2, among others are ANTARA national news agency, e-book and other online articles from various websites. The source of corpus is not in hardcopy but already in digital media. This method is process gathering of corpus which has on file in the form of digital to collect quickly. If not, the process that we need to be longer, every manual document must change to digital document. Validation of digital document needs more time.

1.3. Corpus Processing and Tagging

- Data Processing, source from ANTARA news agency

1. Data source of ANTARA News have domains of sport, science and technology, national news and international news.
2. Process selection to collect needed data.
3. Transformation data process to MSSQL Server 2000.
4. The data is article of Indonesian - English which not yet pairs.
5. The data manually paired by application program.
6. After the article paired, next process is aligned sentence by sentence.
7. Data not ready to use caused still containing punctuation mark.
8. Cleaning punctuation mark from the sentences.
9. Send the sentences to the translator 1 and translator 2 to review.
10. Product of translator1 and translator 2 then review by reviewer.
11. Product of reviewer is ready to be used.



1.4. Corpus Processing and Tagging

- Data Processing, source from ANTARA news agency

1. Data source of ANTARA News have domains of sport, science and technology, national news and international news.
2. Process selection to collect needed data.
3. Transformation data process to MSSQL Server 2000.
4. The data is article of Indonesian - English which not yet pairs.
5. The data manually paired by application program.

6. After the article paired, next process is aligned sentence by sentence.
7. Data not ready to use caused still containing punctuation mark.
8. Cleaning punctuation mark from the sentences.
9. Send the sentences to the translator 1 and translator 2 to review.
10. Product of translator1 and translator 2 then review by reviewer.
11. Product of reviewer is ready to be used.

- Data Process, source from E-Book

1. The data is bilingual article (Indonesian-English) in e-book.
2. The data direct paired per sentence manually.
3. Data not ready to use caused still containing punctuation mark.
4. Cleaning punctuation mark from the sentences.
5. Send the sentences to the translator 1 and translator 2 to review.
6. Product of translator1 and translator 2 then review by reviewer.
7. Product of reviewer is ready to be used.

- Data Process, source from internet (science and technology domain)

1. Science and technology article store in text database.
2. Select the sentence of which do not too long.
3. Cleaning punctuation mark from the sentences.
4. Send the sentences to the translator 1 and translator 2 to review.
5. Product of translator1 and translator 2 then review by reviewer.
6. Product of reviewer is ready to be used.

1.5 SGML corpus tagging

In consideration with a standard Indonesian corpus format, we have look into SGML as the mark up language. SGML played a major part in the BNC project which serve as an interchange medium between the various data-providers, as a target application independent format; and as the vehicle for expression of metadata and linguistic interpretations encoded within the corpus.

From the start of the project, it was recognized that we have to choose a standard format such as SGML in order to maintain the corpus for long term storage and also enable distribution of the data. The importance of SGML as an application independent encoding format is also becoming apparent, as a wide range of applications for it begin to be realized. The basic structural mark up of texts may be summarized as follows. Each of the documents or text source articles making up the corpus is represented by a single <IDDoc> element, containing a header, and a <text> (for written texts) element. The header element contains detailed and richly structured metadata supplying a variety of contextual information about the document (its title, source, encoding, etc., as defined by the Text Encoding Initiative). The headers were automatically generated from information managed within a relational database. A written text is divided into paragraphs, possibly also grouped into hierarchically numbered divisions. Below the level of the paragraph, all texts are composed of <s> elements, marking the automatic linguistic segmentation, and each of these is divided into <w> (word) or <c> (punctuation) elements. At the subsequent processing of the corpus, we expect that POS Tagging performed at University of Indonesia (UI) will insert a POS (part of speech) annotation attribute to each <w> and <c> elements.

1.6 Translating Penn Treebank Corpus into Indonesian

The Penn Treebank Corpus will be used in developing the part-of-speech tagger for Bahasa Indonseia and the machine translation tool. We plan to translate the Penn Treebank Corpus into Bahasa Indonesia manually by several translators from the Faculty of Arts,

University of Indonesia. Then the translated documents will be evaluated by a senior linguist who will review the result of the translation work.

Table 2.1: Penn Treebank and Brown Corpus Tagsets

Penn Treebank Tagset		**Brown Corpus Tagset**	
Tag	**Makna**	**Tag**	**Makna**
NN	Noun	AT	article
NNP	proper noun	BE	be
NNPS	Proper noun plural	CC	coordinating conjunction
NNS	Noun plural	CD	cardinal numberal
RB	Adverb	CS	subordinating conjunction
RBS	Adverb, superlative	HV	have
RBR	Adverb, comparative	IN	preposition
FW	Foreign word	JJ	adjective
IN	Prepositttion/subordinating conjuction	MD	modal auxiliary
CC	Coordinating conjuction	NC	cited word
MD	Modal	NN	singular or mass noun
DT	Determiner	BE	be
EX	Existensial there	CC	coordinating conjunction
JJ	Adjective	CD	cardinal numberal
JJS	Adjective, superlative	CS	subordinating conjunction
JJR	Adjective, comparative	HV	have
POS	Possesive	IN	preposition
TO	To	JJ	adjective
WDT	Wh-Determiner	MD	modal auxiliary
WP	Wh-Pronoun	NC	cited word
WRB	Wh-Adverb	NN	singular or mass noun
WP$	Possesive Wh-Pronoun	NNS	plural noun
PDT	Predeterminer	NP	proper noun
PRP$	Possesive pronoun	OD	ordinal numeral
LS	List maker	QL	qualifier
UH	Interjection	RB	adverb
RP	Particle	RP	adverb/particle
CD	Cardinal Number	TO	infinitive marker to
VB	verb, base form	UH	interjection, exclamation
VBP	verb, non 3rd. singular present	VB	verb, base form
VBN	verb, past tense	VBD	verb, past tense
VBG	verb, gerund	VBG	verb, gerund
VBN	verb, past participle	VBN	verb, past participle
VBZ	verb, 3rd. singular present	VBZ	verb, 3rd. singular present

2. Part-of-Speech Tagger for Bahasa Indonesia

Tagset for a POSTAG is built based on the characteristics and purposes of the POSTAG. Each language usually has its own tagset, however, a language may have several different tagsets. For example, English has Penn Treebank *Tagset* and Brown Corpus *Tagset* (see Table 2.1) [Francis & Kucera, 1982; Marcus et al., 1993].

So far, for Bahasa Indonesia, there is no available POSTAG yet, so we have to base our study on a grammar book [Alwi, 2003]. Our plan is to start with a simple tagset suitable for Bahasa Indonesia based on the Penn Treebank and Brown Corpus Tagsets. We have analyzed some Indonesian articles and chosen to use tags for words that are frequently used in the articles. At this time, we have identified 28 tags that we will consider for Bahasa Indonesia. This is still subject to changes as our study on the POSTAG is still underway.

Table 2.2 Proposed Tagset for Bahasa Indonesia

Category	POS Tag ID No.	POS Name	POS Tag
Noun	*1*	*Countable Common Noun*	*NNC*
	2	*Uncountable Common Noun*	*NNU*
	3	*Genitive Common Noun*	*NNG*
	4	Proper Noun	NNP
Verb	5	Modal	MD
	6	*Transitive Verb*	*VBT*
	7	*Intransitive Verb*	*VBI*
Adjective	8	Adjective	JJ
Adverb	9	Adverb	RB
Preposition	10	Preposition	IN
Conjunction	11	Coordinate Conjunction	CC
	12	Subordinate Conjunction	SC
Pronoun	13	Personal Pronoun	PRP
	14	WH-pronoun	WP
	15	*Number Pronoun*	*PRN*
	16	*Locative Pronoun*	*PRL*
Interjection	17	Interjection	UH
Punctuation	18	Punctuation	PUN
Symbol	19	Symbol	SYM
Determiner	20	Determiner	DT
	21	WH-determiner	WDT
Particle	22	Particle	RP
Cardinal Numeral	*23*	*Primary Numeral*	*CDP*
	24	*Ordinal Numeral*	*CDO*
	25	*Irregular Numeral*	*CDI*
	26	*Collective Numeral*	*CDC*
Negation	*27*	*Negation*	*NEG*
Foreign Word	28	Foreign Word	FW

In order to develop a POSTAG for Bahasa Indonesia, we have studied some methods that are already used on other languages, such as the *Conditional Random Field* [Lafferty et

al., 2001; McCallum, 2003; Wallach, 2004; Avinesh & Karthik, 2007], the *Transformation Based Learning* [Brill, 1992; Brill, 1993; Brill, 1993b; Brill, 1994], the *Maximum Entropy Model* [Ratnaparkhi, 1996], and the Hidden Markov Model [Kupiec, 1992].

Conclusion

This report contains our early work on building language resources and machine translation tool for Bahasa Indonesia. We had done some early work such as literature survey and prepared some plans for building Indonesian corpus by translating Penn Treebank and collecting Indonesian articles. We had also prepared a tagset for developing Indonesian POSTAG for the PAN10L project.

References

1. Alwi, Hasan. et.al. *Tata Bahasa Baku Bahasa Indonesia*, 3rd ed. Balai Pustaka: Jakarta, 2003.
2. Aston, G. and Burnard, L. *The BNC Handbook* Edinburgh: Edinburgh UniversityPress., 1998.
3. Avinesh, P., and Karthik, G. *Part of Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning*. Proceedings of IJ-CAI Workshop on Shallow Parsing for South Asian Languages. 2007.
4. Brill, E. *A Simple Rule-Based Part of Speech Tagger*. Proceedings of the Third Conference on Applied Natural Language Processing. 1992.
5. Brill, E. *Automatic Grammar Induction and Parsing Free Text: A Transformation-Based Approach*. Proceedings of the 31st Meeting of the Association of Computational Linguistics. 1993.
6. Brill, E. *Some Advances in Transformation-Based Part of Speech Tagging*. Proceedings of the 12th National Conference on Artificial Intelligence. 1994.
7. Commons Wikipedia, http://commons.wikimedia.org/wiki/Main_Page. Sept 2008
8. Cutting, D., Kupiec, J., Pederson, J., and Sibun, P. *A Practical Part of speech Tagger*. Proceedings of the Third Conference on Applied Natural Language Processing. 1992.
9. Della Pietra, V. and Lafferty, J. *Including Features of Random Fields*. IEEE Transactions on Pattern Analysis and *Machine* Intelligence, Vol 19, 1997, p. 380–393.
10. Garside, R., Leech, G., and McEnery, T. *Corpus annotation: linguistic information from computer text corpora* Harlow: Addison-Wesley Longman.1997.
11. H Handke, Jürgen. *The Structure of the Lexicon: Human Versus Machine*. Walter deGruyter, 1995.
12. Kupiec, J. M. *Robust Part-of-Speech Tagging Using A Hidden Markov Model*. Computer Speech and Language, Vol 6(3),1992, p. 225-242.
13. Lafferty, J., McCallum, A., and Pereira, F. *Conditional Random Field: Probabilistic Model for Segmenting and Labeling Sequence Data*.. Proceedings of the Eighteenth International Conference on Machine Learning. 2001.
14. Marcus, M., Santorini, B., and Marcinkiewicz, M. *Building A Large Annotated Corpus of English: The Penn Treebank*. Computational Linguistics, Vol 19(2), 1993, p. 313–330.
15. Ratnaparkhi, A. *A Maximum Entropy Model for Part-of Speech Tagging*. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Vol.1996, p. 133–142.
16. Sharof, Serge. *In the garden and in the jungle: comparing genres in the BNC and Internet*. http://corpus.leeds.ac.uk/serge/webgenres/colloquium/, Sept 2008.
17. Wikipedia Creative Commons, http://en.wikipedia.org/wiki/, retrieved August 08