IDENTIC Corpus: Morphologically Enriched Indonesian - English Parallel Corpus

Septina Dian Larasati

Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics,
Prague, Czech Republic
SIA TILDE
Riga, Latvia
larasati@ufal.mff.cuni.cz, septina@tilde.lv

Abstract

This paper describes the creation process of an Indonesian-English parallel corpus (IDENTIC). The corpus contains 45,000 sentences collected from different sources in different genres. Several manual text preprocessing tasks, such as alignment and spelling correction, are applied to the corpus to assure its quality. We also apply language specific text processing such as tokenization on both sides and clitic normalization on the Indonesian side. The corpus is available in two different formats: 'plain', stored in text format and 'morphologically enriched', stored in CoNLL format. Some parts of the corpus are publicly available at the IDENTIC homepage.

Keywords: Indonesian, Corpus, Morphology

1. Introduction

Building a language resource is one of the main and the earliest stages in Natural Language Processing (NLP) research. Having a proper language resource is one of the main challenges for an under-resourced language. Since in most cases, the language resources are independently compiled and processed in a small project or research group and rarely shared. While for well researched languages, such as English, German, or Czech, there are plenty of language resources to work on which can be referred to.

Indonesian, or Bahasa Indonesia as the locals would call it, is one of the most frequently spoken languages in the world due to the country's large population. It is spoken by approximately 230 million speakers which includes its 30 million native speakers. In spite of that fact, the NLP research for this language is not so prolific. Most of the research on textual data-driven methods do not have any proper textual data set to apply their research methods and compare their research outcomes againts.

This paper describes a corpus creation process of an under resourced language, Indonesian. The corpus is a bilingual corpus paired with English. The aim of this work is to build and provide researchers a proper Indonesian-English textual data set and also to promote research in this language pair. This corpus is referred as 'IDENTIC' and available at its homepage[1]. The corpus contains texts coming from different sources in different genres. This work includes three parts. Those parts are manual preprocessing, text processing, and automatically enriching the corpus with morphological information. The corpus is now publicly available in two different formats: 'plain', stored in text format and 'morphologically enriched', stored in CoNLL format (Buchholz and Marsi, 2006) as a flat tree without any dependency construction. Most of the work is focused on the Indonesian side of the corpus.

[1] http://ufal.mff.cuni.cz/ larasati/IDENTIC.html

2. Indonesian Language Properties

Indonesian, one of the Austronesian languages, uses the Latin alphabet with 26 letters, which makes the corpus easily stored without any special encodings. The language is not an inflectional language such as Slavic or Baltic languages that changes the word forms depending on the case or gender, but it has many word derivational cases. Indonesian has a strict SVO word order similar to English, but it has a different phrasal head-modifier order. Here are listed several specific Indonesian language properties that we encountered and handled during this work.

Reduplication Indonesian uses reduplication to mark plurality of the word. This not only applies to Noun Phrases but also to Verbs and Adjectives. On Verbs, the reduplication marks events that are done several times or habitual. On Adjectives, the reduplication conveys the reference of the Adjectives' nature belonging to plural entities. The reduplicated words are separated by a hyphen e.g. *'kucing-kucing'* which means 'cats'.

Clitic Several Personal Pronouns can be put as a separate word and also can be formed as clitics glued to the Verbs or Noun Phrases. Those clitics then become the participants of the Verb events or a Possessive Pronoun of the Noun Phrase e.g. "*kupeluk kucingku*" *(ku+peluk kucing+ku)* which means "I hug my cat".

3. Data Sources

The corpus contains texts coming from different sources which makes its sentences vary in genres and language styles. Some part of the corpus are taken from PAN localization project (BPPT, 2010) output, which mostly contains articles in formal language style. Some other parts of the

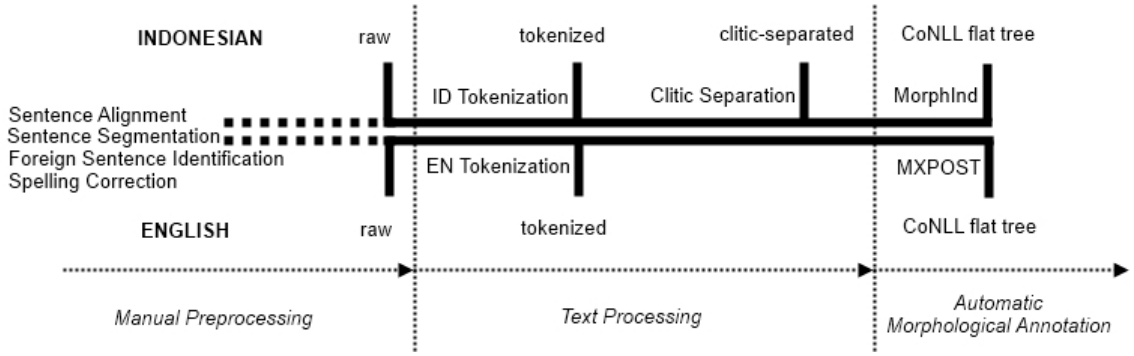

Figure 1: The Corpus Creation Stages.

corpus are taken from a free subtitle provider site [2] which mostly contains spoken dialogue sentences. There are also a few comparable sentences taken from several news sites, which are not the exact translation of one another. The text sources of the corpus can be seen in Table 1.

Source	Description
PC,PS, PI,PE	Indonesian corpus with its parallel English translation in four different genres. Those genres are Science (PC), Sport (PS), International (PI), and Economy (PE), taken from PAN Localization Project.
PP	Indonesian corpus of translated Penn Treebank (Marcus et al., 1993) sentences, which are also taken from PAN Localization project. The English side is not provided by PAN, but provided by Linguistic Data Consortium (LDC).
NW	Manually downloaded comparable articles from the several news websites.
SB	Manually downloaded movie subtitles.

Table 1: Data Source

4. Corpus Creation Stages

The corpus creation is done in three stages, namely *Manual Preprocessing*, *Text Processing*, and *Automatic Morphological Annotation*. The general schema of the corpus creation stages can be seen in Figure 1.

4.1. Manual Text Preprocessing

Several manual text preprocessing tasks are applied to the source texts and those tasks are as follows:

SA (Sentence Alignment): identifying the pairs of sentences in different languages that are translation of each another or convey the same meaning.

SS (Sentence Segmentation): identifying the sentence boundaries.

SFI (Foreign Sentence Identification): identifying the foreign or untranslated sentences in the text. All the foreign sentences are deleted.

[2] http://www.opensubtitles.org/

SC (Spelling Correction): correcting the misspelled words.

Most of the texts that come in parallel sentences, as it is taken from the source (i.e. **PC,PS,PI,PE**), are properly aligned. Most manual preprocessing tasks are done for **PP, NW**, and **SB** text, since those texts are not aligned. And in the case of **PP**, PAN does not provide the English side because its license belongs to the LDC. The **PP** sentences on the Indonesian side are segmented according to the English side. This is done to keep the reference to the PENN Treebank syntactic structure for later research if needed. The **PP** English side is not publicly available at the IDENTIC homepage.

	#Sentences (ID-EN)		Preprocessing			
	Before	After	SA	SS	FSI	SC
PC	6,355	6,355			●	●
PS	4,483	4,465			●	●
PI	6,644	6,641			●	●
PE	6,540	6,540			●	●
PP∗	23,468	17,674	●	●	●	●
NW∗∗	164	164	●	●	●	●
SB∗∗	3,161	3,161	●	●	●	●
Total		45,000				

Table 2: Corpus statistic of the number of sentences and the manual text preprocessing tasks applied.
∗) Only Indonesian sentences are provided by the source.
∗∗) The text is not (properly) segmented from the source.

4.2. Text Processing

The text processing tasks applied are *Tokenization* and *Clitic Normalization*. The Tokenization is applied on both languages differently. Although both use same general tokenizer tool, we add several additional language specific rules. The tokenizer used in this text processing is MOSES's (Koehn et al., 2007) tokenizer script v.3. The language specific rules are explained as follows:

4.2.1. Tokenization:English

An additional tokenizer rule is added to separate the token in the pattern of *'rp'* followed by a nominal, which conveys Indonesian currency e.g. *'rp7000'* to become *'rp 7000'*.

Indonesian

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS
1	ku	aku	aku<p>_PS1	PS1	p|P|S|1|aku|1|1
2	mencintai	cinta	meN+cinta<n>+i_VSA	VSA	n|V|S|A|meN+cinta+i|1|1
3	mu	kamu	kamu<p>_PS2	PS2	p|P|S|2|kamu|1|1
4	.	.	.<z>_Z–	Z–	z|Z|-|-|.|1|1

English

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS
1	I	I	I_PRP	PRP	PRP
2	love	love	love_VBP	VBP	VBP
3	you	you	you_PRP	PRP	PRP
4	.	.	.	.	.

Figure 2: The snippets of IDENTIC 'morphologically enriched' type, stored in 2006 CoNLL Shared Task Data Format. The fields HEAD, DEPREL, PHEAD, and PDEPREL are omitted since the values will always be set to '0', 'ROOT', '_', and '_' respectively

4.2.2. Tokenization:Indonesian

For Indonesian sentences, we add some rules to handle hyphenated word forms. Hyphenated words are usually separated in the tokenization, but that is not a general rule in Indonesian. Reduplicated words are constructed with a hyphen and this construction should not be separated. In this case, we use MorphInd, an Indonesian Morphological Tool[3] (Larasati et al., 2011) to detect whether a surface word is reduplicated or not. With MorphInd, the reduplicated word is analyzed as plural form while the others are separated by a 'DASH' marker, as shown in a simple example in Figure 3.

(1) *'kucing-kucing'* (cats):
ˆkucing<n>_NPD$
is kept in *'kucing-kucing'* form

(2) *'melambai-lambai'* (waving repeatedly):
ˆmeN+lambai<v>_VPA$
is kept in *'melambai-lambai'* form

(3) *'amerika-jepang'* (america-japan):
ˆamerika<n>_NSD$DASHˆjepang<n>_NSD$
is changed into *'amerika - jepang'* form

Figure 3: The plural forms for Nouns (1) or Verbs (2) are marked as plural, while the hyphenated multi words (3) are marked with a 'DASH' marker.

4.2.3. Clitic Normalization:Indonesian

The Clitic Normalization is only done on the Indonesian side. This is done to normalize the word forms into their independent lexical unit. This is also done automatically by looking at the MorphInd output which has clitic markings (as seen in Figure 4). With that information we separate the necessary clitic(s) from the main word.

[3]http://ufal.mff.cuni.cz/ larasati/MorphInd.html

'kumencintaimu' (I love you):
ˆaku<p>_PS1+meN+cinta<n>+i_VSA+kamu<p>_PS2$

Figure 4: Clitic Normalization. The word will be separated as ∗*"ku mencintai mu"*, which is not the correct form. The correct form is *"aku mencintai kamu"*.

4.3. Automatic Morphological Annotation

The morphological annotation is done automatically. It is stored in 2006 CoNLL Shared Task format because of its simplicity as to compare to XML and to accommodate future research on Indonesian dependency parsing. Currently the corpus is stored without any dependency information. The HEAD field is filled with the value '0' which points to the root node and the DEPREL field's value is 'ROOT'. The morphological annotation on both sides is described as follows:

4.3.1. English: using MXPOST

The morphological information on the English side contains the part-of-speech tag provided by MXPOST tagger (Ratnaparkhi, 1996).

4.3.2. Indonesian: using MorphInd

We also use MorphInd to provide the morphological information on the Indonesian side. MorphInd is chosen for its broader coverage and detailed analysis. Compared to the other Indonesian morphological analyzer (Pisceldo et al., 2008), MorphInd has broader coverage of 84.69±0.28% as to compare to 81.91±0.18%. MorphInd analysis also has a richer tagset and gives morphemic segmentation information including clitics.

MorphInd analysis fills the LEMMA, CPOSTAG, POSTAG, and FEATS fields in CoNLL Format. Clitics are treated as individuals words, but marked as clitics. The FEATS field consists of more detailed morphological information and it is filled with the following seven different features:

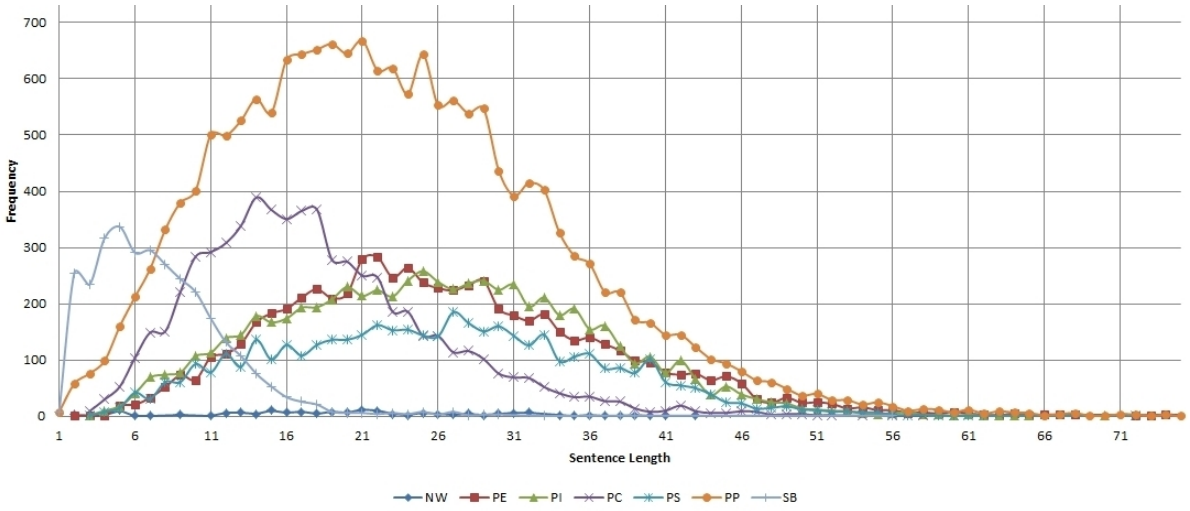

Figure 5: Indonesian sentence length frequency.

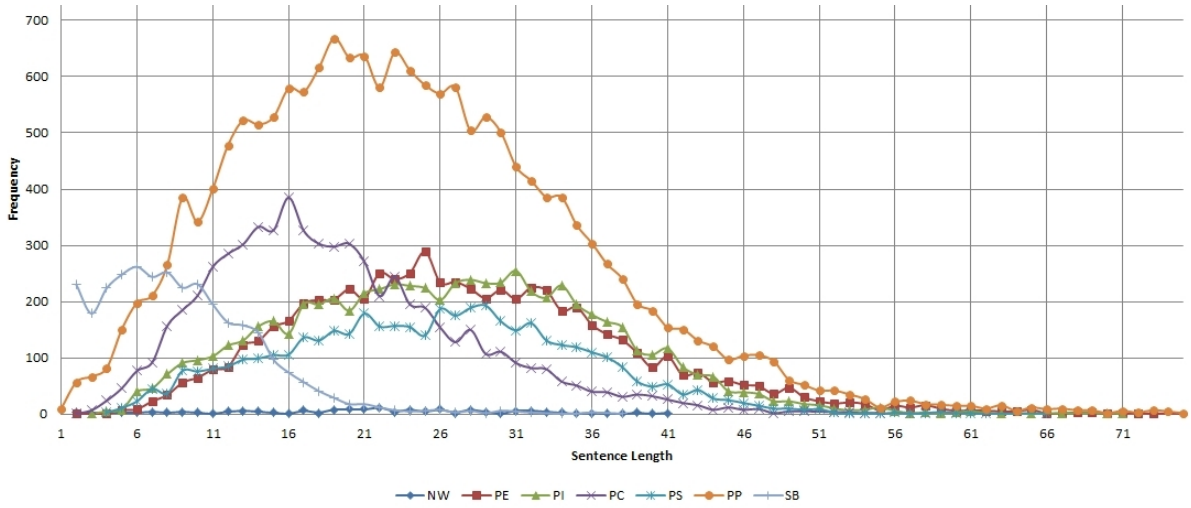

Figure 6: English sentence length frequency.

1. MorphInd Lemma tag.
2. MorphInd morphological tag position 1.
3. MorphInd morphological tag position 2.
4. MorphInd morphological tag position 3.
5. MorphInd morphemic segmentation.
6. Flag for 'no spacing before'. It applies for clitic and punctuation. Filled with '0' if there is a spacing before, and '1' if there is no spacing before.
7. The word ID to which it is glued to (for clitics and punctuation cases with 'no spacing before').

Given in Figure 2 is the corpus snippet example for the Indonesian sentence *'Kumencintaimu'* which is analyzed by MorphInd as shown in Figure 4.

5. Downloadable Resources

We provide the corpus in three kinds of the 'plain' type for the Indonesian side (*raw*, *tokenized*, and *clitic-separated*) and two for English (*raw* and *tokenized*). The 'plain' type is stored in the following text format as seen in Figure 7. The 'plain' type snippet can be seen in Figure 8.

FORMAT: [ID][tab][id sentence][tab][en sentence]

Figure 7: The 'plain' type text format.

(1)	subtitle-...	Kumencintaimu.	I love you.
(2)	subtitle-...	Kumencintaimu .	I love you .
(3)	subtitle-...	Ku mencintai mu .	I love you .

Figure 8: IDENTIC 'plain' type snippet stored in text format.

The snippets for the 'morphologically enriched' type can be found in Figure 2.

6. The Corpus Statistics

The corpus has in total of 45,000 sentences. Since the sentences are coming from different genres and having dif-

ferent styles, they also differ in sentence length. Given in Figure 5 and 6, is the sentence length frequency in Indonesian and English. Most of the sentences coming from subtitles are short sentences. Sentences coming from articles have similar sentence length distribution among themselves. Given in Table 3 are some others statistics of the corpus:

	ID		EN	
	#words	vocabulary size	#words	vocabulary size
PC	110,996	11,402	123,333	11,896
PS	112,053	8,232	118,682	9,386
PI	167,703	11,776	178,974	13,683
PE	168,775	11,761	185,109	11,304
PP	407,517	23,263	435,265	25,296
NW	3,208	1,221	3,608	1,286
SB	24,293	3,114	29,769	2,938
Total	994,545		1,074,740	

Table 3: IDENTIC number of words and vocabulary size statistics.

7. Future Work

IDENTIC is open for any free available texts that want to be compiled together as IDENTIC. The new added corpus will at least be preprocessed similar to the work described here. There are many other annotations that can be added to IDENTIC, such as syntactic annoatation (either constituency or dependency approach), named entity, terminology annotation, etc. Manual annotation can also be applied to IDENTIC to create a gold-standard corpus for any type of research.

8. Conclusion

Parts of the work results described in this paper are publicly available and follows the source text licenses. The corpus contains 45,000 parallel sentences in Indonesian and English and is available (except sentences that are part of PENN Treebank) in two different formats: 'plain' and 'morphologically enriched'. The corpus can be found at the IDENTIC homepage.

9. Acknowledgment

The research leading to these results has received funding from the European Commission's 7th Framework Program under grant agreement n° 238405 (CLARA), by the grant LC536 Centrum Komputační Lingvistiky of the Czech Ministry of Education, and this work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

10. References

BPPT, 2010. *Research Report on Corpus Design and Collection and Cleaning Tools English to Bahasa Indonesia.*

S. Buchholz and E. Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics.

J. Hajič. 2004. Complex corpus annotation: The prague dependency treebank. *Insight into the Slovak and Czech Corpus Linguistics*, page 54.

L.S. Indradjaja and S. Bressan. 2003. Automatic learning of stemming rules for the indonesian language. In *Proc. of the The 17th Pacific Asia Conference on Language, Information and Computation (PACLIC)*.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

S.D. Larasati, V. Kuboň, and D. Zeman. 2011. Indonesian morphology tool (morphind): Towards an indonesian corpus. *Systems and Frameworks for Computational Morphology*, pages 119–129.

M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

B. Nazief. 2000. Development of computational linguistics research: A challenge for indonesia. *Computer Science Center, University of Indonesia*.

P. Pajas and J. Štěpánek. 2008. Recent advances in a feature-rich framework for treebank annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 673–680. Association for Computational Linguistics.

F. Pisceldo, R. Mahendra, R. Manurung, and I.W. Arka. 2008. A two-level morphological analyser for the indonesian language. In *Australasian Language Technology Association Workshop 2008*, volume 6, pages 142–150.

F. Pisceldo, M. Adriani, and R. Manurung. 2009. Probabilistic part of speech tagging for bahasa indonesia. In *Third International MALINDO Workshop, co-located event ACL-IJCNLP*.

A. Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. University of Pennsylvania.

H. Riza. 2008. Resources report on languages of indonesia. *IJCNLP 2008*, page 93.