PAN Localization Project

RESEARCH REPORT

PHASE 1.1

Initial Research Report on
Corpus Design, Collection and Cleaning Tools

Center for Science and Technology Information Network(IPTEKnet)

Agency for the Assessment and Application of Technology (BPPT)

1. Design of Corpus

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 (see: The BNC Handbook: Exploring the British National Corpus with SARA, by Guy Aston and Lou Burnard, Edinburgh University Press, 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 hyper linking, 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 Serge Sharof "In the garden and in the jungle: comparing genres in the BNC and Internet" (see http://corpus.leeds.ac.uk/serge/webgenres/colloquium/), 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. In this presentation I will 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).

Based on the above design consideration, 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

2. 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)**

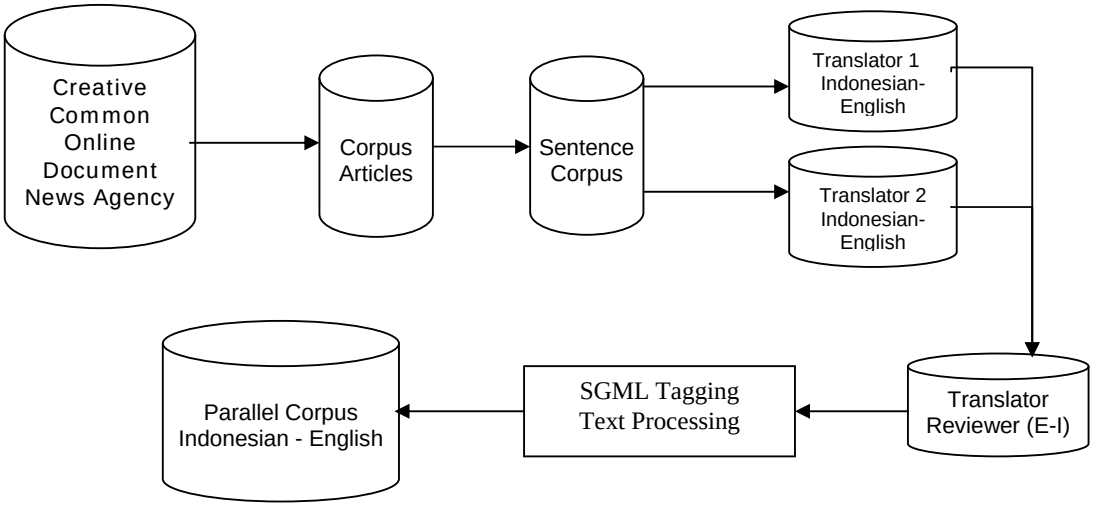
Table 2. Corpus Sources Contact Information

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

3. 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.



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.

Corpus tagging

In consideration with a standard Indonesian corpus format, we have look into SGML/XML 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. Text Encoding Initiative (TEI) uses XML as part of the encoding of documents. From the start of the project, it was recognized that we have to choose a standard format such as SGML/XML in order to maintain the corpus for long term storage and also enable distribution of the data.

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.

5. Schedule

The following schedule is used for doing our work during the first 2 phases, Phase 1.1 (July-Sept 08) and Phase 1.2 (Oct-Dec 08).

Task Name for Phase 1.1 & Phase 1.2	Jul	Aug	Sep	Oct	Nov	Dec
Corpus Design and Collection Plan						
Corpus Source Identification						
Collection Trial						
Corpus Collection 100K words Indo						
Translation of 100K words (Indo-Eng)						
Corpus Processing 100K words						
Design of Machine Translation						

6. Team

The PANL10n team at BPPT comprises of the following members:

1. Hammam Riza (Group Leader)
2. T. Syarfina (Linguist)
3. Arief Sartono (Translator Reviewer)
4. Eggi Githa Nagara (Translator)
5. Awal Subandar (Translator)
6. Darwito (Translator)
7. Budiono (Research Officer)
8. Henky Mulyadi (Research Officer)
9. Adiansya Prasetya (Research Officer)

The team has a tied schedule to produce 1000 sentences each week to put into the corpus development system, in order to achieve the targeted 100K words (approximately 10K sentences). It will take around 10 weeks of corpus collection, translation and processing to meet the goal of Planned Work on Phase 1.2.

7. Reference

1. Jürgen Handke, The Structure of the Lexicon: Human Versus Machine, Walter de Gruyter, 1995
2. Wikipedia Creative Commons, http://en.wikipedia.org/wiki/, retrieved August 08
3. Commons Wikipedia, http://commons.wikimedia.org/wiki/Main_Page, Sept 2008
4. Aston, G. and Burnard, L. *The BNC Handbook* Edinburgh: Edinburgh University Press., 1998
5. Serge Sharof "In the garden and in the jungle: comparing genres in the BNC and Internet, http://corpus.leeds.ac.uk/serge/webgenres/colloquium/, Sept 2008
6. Garside, R., Leech, G., and McEnery, T. *Corpus annotation: linguistic information from computer text corpora*, Harlow: Addison-Wesley Longman, 1997.