

Machine Translation for Indonesian and Tagalog

Brianna Laugher

ToggleText
Melbourne, 3191 Australia
briannalaugher@toggletext.com

Ben MacLeod

ToggleText
Melbourne, 3191 Australia
benmacleod@toggletext.com

Abstract

Kataku is a hybrid MT system for Indonesian to English and English to Indonesian translation, available on Windows, Linux and web-based platforms. This paper briefly presents the technical background to Kataku, some of its use cases and extensions.

Kataku is the flagship product of ToggleText, a language technology company based in Melbourne, Australia.

1 Background to Indonesian to English MT

1.1 The Indonesian language

Indonesian is a variety of the Malay language. Both Indonesian and Tagalog belong to the Malayo-Polynesian subgroup of Austronesian languages, although Tagalog is a Central Philippine language under the Borneo-Philippine languages branch, and Indonesian is a Sunda-Sulawesi language under the Nuclear Malayo-Polynesian branch. (Wouk & Ross, 2002) Indonesian and Tagalog are perhaps as similar as English and Latin -- "distant cousins" at best.

Indonesian has a reputation as an easily learnt language for English speakers, with a basic SVO word order and limited morphology. There is no grammatical gender, and tense/aspect is marked by time adverbs and aspectual particles. Its writing system is generally regular and almost all of its phonemes are familiar to English speakers. There are relatively few brand new concepts for English

monolinguals, for example, the use of classifiers (measure words).

1.2 The country of Indonesia

Indonesia has a population of over 220 million people, making it the world's fourth most populous country and the most populous Muslim nation (over 85% of the population is Muslim), although it is not an Islamic state. It is comprised of thousands of islands north of Australia and south of Singapore, Malaysia and the Philippines. Papua New Guinea lies to the east and it shares an island with that country, as well as Malaysia and Timor-Leste (East Timor).

The local variety of Malay became the official language of Indonesia (*Bahasa Indonesia*) in the 1920s while Indonesia was under Dutch rule. At that time it was the language of trade and inter-region communication.

Regional languages are still frequently used in informal settings, especially in rural areas. Although there are other languages in Indonesia with larger numbers of native speakers (especially Javanese), Indonesian is widely spoken as a second language. After the first three years of primary school, all education is conducted in Indonesian. It is also used by the government and the mass media. (Worsley, 1994)

1.3 Technology survey

There are no tagged or parallel corpora available for Indonesian, which greatly limits the potential for statistical MT systems to be functional.

Aside from Kataku, there are very few tools, web-based or otherwise, available for translating to or from Indonesian.

*Transtool*¹ is Windows-based software that performs bi-directional MT and claims a lexicon of over 200,000 entries. It was released in 2004 and is now reported to be widely available among Indonesian PCs thanks to software piracy. In our experience it has been unreliable with semi-regular crashes.

*ReksoTranslator*² is similar such software claiming 216,000 English to Indonesian entries and 150,000 Indonesian to English.

LEC³ offers several products including *Translate SDK*, *Translate DotNet* and *Translate2Go* which translates multiple language pairs, including Indonesian to English. These appear to be “shell” products that merely communicate with LEC’s translation servers via the internet, rather than conducting translations on a local install.

1.4 Translation comparisons

Much work remains to be done in evaluating and comparing MT output, but as a very basic comparison we translated this text, from Wikipedia⁴, across several systems:

Sejarah Indonesia banyak dipengaruhi oleh bangsa lainnya. Kepulauan Indonesia menjadi wilayah perdagangan penting setidaknya sejak sejak abad ke-7, yaitu ketika Kerajaan Sriwijaya menjalin hubungan agama dan perdagangan dengan Tiongkok dan India.

LEC⁵: *A history Indonesia much dipengaruhi by a nation lainnya . An archipelago Indonesia a domain becomes perdagangan shoot away a poem a poem a century to - 7 , yaitu when a kingdom South Sumatran Maritime Kingdom tie together a religion and perdagangan China and India .*

Transtool⁶: *The history of Indonesia Indonesian history influenced many by other nation. Archipelago of Indonesia become important commerce region at least since since century of ke-7, that is when Empire of Sriwijaya braid religion [relation/link] and commerce with Tiongkok and India.*

ToggleText's Katak⁷: *The Indonesian history often was influenced by the other nation. The In-*

doneasian island became the important trade territory at least since since the 7th age, that is when the Sriwijaya Kingdom established religious relations and the trade with Tiongkok and India.

ReksoTranslator was unavailable for this comparison.

2 System presentation

2.1 Architecture and requirements

Katak⁷ is web-based and compatible with most platforms. It has a high speed performance, translating up to 1200 to 2400 words per minute. The system architecture is modularised to provide the best possible leverage and permit the re-use of modules for expanding into other language pairs. The system is developed and by default runs on the Linux operating system. Katak⁷ is developed using test-driven development methodology, ensuring a reliable and consistent product.

Katak⁷ is a hybrid MT system. It contains elements of a traditional rule-driven MT architecture, but uses Prolog's powerful logic programming to imbue such rules with context sensitivity. The predicate (verb) “rules” are driven by the rich lexicon. This is necessary because the distance between the language pair (Indonesian and English) is so vast. The lexicon is large and general, covering most common domains.

Its “failsafe” module is built for robustness over messy or ill-formed input. It has been built as a product, meaning development has not been overly concentrated on just parsing.

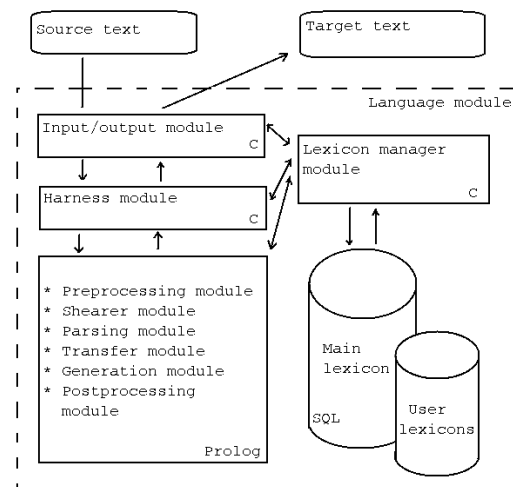


Figure 1: System diagram.

¹<http://www.geocities.com/cdpenerjemah/>

²<http://rekso-inovasi.com/>

³<http://www.lec.com/>

⁴<http://id.wikipedia.org/wiki/Indonesia>

⁵<http://www.lec.com/translate-demos.asp>

⁶Via a purchased copy of this software.

⁷<http://toggletext.com/main.cgi?page=translation>

2.2 Access and delivery modes

Since Katakku was developed it has been available on the ToggleText website⁸, allowing translation of up to 300 words of free text or web page content. The website currently receives over 70,000 visitors per month.

For our more security conscious customers **Katakku Enterprise** has been available. Katakku Enterprise is a Windows package that allows the software to be installed locally, avoiding the need for internet access. This also comes with the Lexicon Editor. Katakku Enterprise is also available on Linux, providing access via both a GUI and a flexible command-line API. This product is not available to the general public to avoid risking software piracy.

Katakku WebCS provides business and other multi-user organizations with a web interface for bi-directional translation on their own intranet. This intranet interface funnels requests from the organisation across the internet to a ToggleText translation engine. Through this intranet interface, the organisation's users can make translation requests of text documents, free text and web pages, as well as perform dictionary searches. Customers can change the interface's 'look and feel' to blend in with their own intranet. The product also allows the organization's network administrators to monitor the translation server's utilization and performance.

It is anticipated that **website accounts** will soon be available, allowing for individuals to purchase larger word and volume limits for their use of Katakku's web-based interface.

2.3 Strengths

Katakku is particularly strong in handling time phrases and proper noun phrases. It has been designed for robust handling of badly structured input. Its modular design allows for flexible extensibility, for example with the incorporation of statistical methods.

2.4 Limitations

Katakku has not had specialised (i.e., domain-specific) lexicons developed for its default lexicon yet, although ToggleText is available for such

work (and such lexicons may be added by Katakku Enterprise users via the Lexicon Editor).

3 Use case studies

ToggleText has experienced a variety of government users, including Australian, French, American and Canadian clients.

With the growing field of information analysis and retrieval tools, MT has now been given a new life in that it is now being introduced as just one more tool within an integrated suite of larger software tools which delve for scattered, non-standardised information. This push, and acceptance, of MT in Australia has come from two areas. The first is the Defence Science and Technology Organisation (DSTO), an Australian governmental research organisation, whose role it is to monitor and assess new technology for eventual introduction into government departments. This group has been a great support, demonstrating to government users the usefulness of MT when it is an integral part of a larger information retrieval situation.

Similarly, Katakku has been presented within MITRE's Clipper and Trim programs to good effect in the USA. This shift away from pure translation assistance, to integration within a larger suite of information analysis tools is the single biggest change we have witnessed in both government, NGO and corporate groups.

The other type of support we have received is from governmental groups who are in the field, for example, in cooperative work with the police and army of other nations in the arenas of international drug smuggling, terrorism, and other serious crime detection and resolution. In other words, when people have an immediate need for an MT system, then they will accept it willingly and use it effectively.

Katakku is also used in an international effort to track outbreaks of infectious diseases in areas which are remote and otherwise poorly tracked by local officials. This again requires a wide net of information retrieval to look for indications of outbreaks, rather than reliance upon formal official reports.

The tsunami that hit Aceh province in Indonesia killed hundreds of thousands and left most of the country devastated. Acehnese dialect is very similar to formal Indonesian, differing mostly by a phonological shift of vowels, so that ToggleText could easily have adapted Katakku to translate Ace-

⁸<http://www.toggletext.com/>

hinese. However, our attempts to offer support in this way were not successful, because there was no available delivery mechanism, no mobile networks, no electricity even. And yet, in the future, this wider area of international aid work will come into its own.

3.1 Commercial users

One of our recent commercial users is a large resources company based in Indonesia, utilising the WebCS product. They have used it to conduct over 60,000 translations each month, approximately 1.4 million words (this includes dictionary usage). The WebCS product has proved to be reliable and robust, with no downtime.

4 Malay

4.1 Malay

We have received reports of Katakaku being used to translate Malay, although we do not recommend this.

5 Tools and extensions

5.1 Lexicon editor

The Katakaku Enterprise Lexicon Editor is a flexible and powerful Windows application for Katakaku Enterprise that allows users to enter new words into the lexicon or change the translation of existing entries. In a multi-user environment, each user can define their own lexicon (or in fact multiple lexicons), as well as set read/write permissions for other users. Users can then create “super-lexicons” and specify multiple lexicons to be called upon and their priority.

The default lexicon is not directly editable by users, and closed-class parts of speech are not able to be added via the Lexicon Editor. This protects the system from permanent damage by otherwise well-meaning users.

5.2 API

The Linux command-line API allows translation

by specifying an input source text file and an output file to write to. A variety of translation modes may also be optionally specified, that adjust the output format accordingly. For example, it is possible to specify “tagging mode”, which outputs a part of speech tagged file instead of a translation. It is also possible to specify particular user lexicons to be used (as described above), or SMS mode (below).

Use of the API interface is recommended for high-volume throughput.

5.3 SMS

Katakaku has an “email/SMS” mode. This is available by ticking a check box on the web-based translation or specifying an API option. Via partnerships with third-party providers, mobile phone users can receive instant SMS translations by sending a text message (with the content to be translated) to a particular phone number. This is then forwarded to our servers, translated under SMS mode and sent back. This service is currently available in Indonesia and Japan.

This mode was not merely a matter of adding slang to the lexicon, although abbreviations, misspellings and dialect slang were definitely major components of the additional lexicon required. “Personal erotic literature” was found to constitute a large portion of SMS use and this also influenced the necessary additional vocabulary. It was also necessary to recognise words with dropped morphology and “loosen” the parser accordingly. It was even necessary to build a language identifier as an initial step, as it was found users just didn't follow the instructions to specify which translation direction they desired.

6 A brief word on Tagalog

Development is currently underway on a Tagalog to English MT system. While we had hoped to be able to leverage a great deal of code from Katakaku, based on the language similarities between Indonesian and Tagalog, we soon realised the linguistic distance was further than we had anticipated. Tagalog features:

- what is surely one of the world's richest morphology systems, with prefixes, infixes, suffixes and assimilation/lenition
- a complex focus system that uses verb morphology to mark the relationship between a

verb and its noun phrases, allowing for phrase movement within a sentence

- a liberal dosing of English words and phrases sprinkled haphazardly throughout any given text (English, along with Tagalog, is an official language and most Filipinos are at least bilingual; furthermore, unlike Indonesian, Tagalog has not benefited from an official decree against the infiltration of English).

The Tagalog to English system will no doubt be much less polished than Katakau, but its completion will surely be a more amazing feat. It is expected to be completed in the second half of 2009.

7 Company background

ToggleText was established 1993 in Melbourne, Australia by Helen McKay, the current Company Director, and Adrienne Osbourne. It was originally established as a translation and localisation service bureau. ToggleText is now a contractor to the Australian government.

In 1998, ToggleText expanded to include a Natural Language Engineering business section, with a focus on the creation and delivery of machine translation systems for the Internet, specialising in translations between Asian languages and English.

As well as MT, ToggleText has also created a standards-compliant, web-based multilingual terminology management solution, designed to target the needs of government linguists and terminology users, currently in the final stages of development.

The intellectual property rights of ToggleText's machine translation products are totally owned by ToggleText Pty Ltd.

Acknowledgments

Thanks to Helen McKay and Mike Dillinger for their assistance.

References

- Worsley, Peter. 1994. *Unlocking Australia's Language Potential. Profiles of 9 Key Languages in Australia. Volume 5 Indonesian/Malay*. National Languages and Literacy Institute of Australia, Canberra.
- Wouk, Fay and Malcolm Ross (eds). 2002. *The history and typology of western Austronesian voice systems*. Pacific Linguistics. Canberra: Australian National University.