

THE ELEVENTH
INTERNATIONAL SYMPOSIUM ON MALAY/INDONESIAN LINGUISTICS

6-8 August 2007
Manokwari, Indonesia

Abstract

This paper describes our proposed project of building a balanced and representative corpus of Indonesian as part of a bigger project of developing a machine-usable grammar within the LFG-based framework of ParGram (Parallel Grammar Project). We discuss the aims of our Indonesian Pargram Project, and why we need to develop our own corpus for the project. We also discuss sources of corpus materials, corpus design and requirements, and challenges in our project implementation.

Challenges of developing a balanced and representative corpus for Indonesian ParGram

THIS DRAFT 05/07/07

I Wayan Arka (ANU), Jane Simpson (Sydney U),
Avery Andrews (ANU), Mary Dalrymple (Oxford U)

1 Introduction

Indonesian is one of the world's major spoken languages with around 190 million speakers (BPS 2004). However, there is little computational linguistic work on the language. This work is needed for automating many natural language processing (NLP) tasks.

This paper outlines our proposed project of building a balanced electronic corpus of Indonesian as part of a bigger project of developing a large-scale machine-usable grammar and lexicon of Indonesian within the LFG-based framework of ParGram (Parallel Grammar Project). We will refer to our project as I-ParGram (Indonesian Parallel Grammar) for short.

The objective of I-ParGram is to develop a good computationally tractable lexicon and grammar for modern Indonesian. It has the potential to produce two distinct related sets of results. One is an empirically robust and linguistically well-grounded analysis of Indonesian. I-Pargram involves rigorous analysis and testing, and will therefore produce a comprehensive description of Indonesian with more in depth analysis of the grammar and lexicon than is currently available. The other is to produce a robust and sophisticated electronic data base of Indonesian language (richly annotated texts and lexicon) with a grammatical parser, building on some preliminary work.

While our corpus development is part of LFG-based I-ParGram Project, the corpus will be designed with different types of uses and users in mind (see §3.4 below).

The paper is structured as follows. We provide an overview of ParGram in §2, followed by the description of the aims of I-ParGram §3.1, the significance of Indonesian Corpus §3.2, sources of corpus materials §3.3, and corpus design §3.4. Finally, main challenges in implementation and collaborative research are given in §4.

2 ParGram: an overview

The Parallel Grammar project (<http://www2.parc.com/istl/groups/nltt/Pargram/>) is an international collaborative research project for the development of large-scale computationally tractable grammars and lexicons of the world's (major) languages. Members include the corporate research laboratories and universities: the Palo Alto Research Center (PARC), Powerset, and Stanford University (USA); Oxford University, Manchester University and the University of Essex (UK), Fuji Xerox (Japan); the Universities of Stuttgart and Konstanz, (Germany), the University of Bergen (Norway),

Sabancı University (Turkey), University of Debrecen (Hungary), and Ho Chi Minh Institute of Information Technology (Vietnam).

Current Pargram analyses (Butt et al. 2002) are the result of over fourteen years of research and discussion (started in 1994), based on data from a typologically wide range of languages (English, German, French, Japanese, Norwegian, Urdu, Welsh and Malagasy). The approach is to develop and process grammars in parallel. Similar analyses and similar technical solutions, wherever possible, are given for similar structures in each language. Parallelism has the computational advantage that the grammars can be used in similar applications and that machine translation (Frank 1999) can be simplified. However, Pargram also allows flexibility where parallelism is not maintained when different analyses are desirable and justified for good language-specific reasons. An encouraging result from Pargram work is the ability to bundle grammar writing techniques into transferable knowledge and technology from one language to another, which means that new grammars can be bootstrapped in a relatively short amount of time (Kim et al. 2003).

The underlying syntactic framework for Pargram is LFG (Lexical-Functional Grammar), a stable and mathematically well-understood constraint-based theory of linguistic structure (Bresnan 2001; Dalrymple 2001; Kaplan and Bresnan 1982). Two important structures are assumed in LFG, constituent structure (*c-structure*) and functional structure (*f-structure*). The *c-structure* representation captures surface (overt) linguistic expressions that vary across languages. It is modelled in phrase structure trees that show structural dominance and precedence relations of units. *F-structure* captures abstract relations of predicate argument structures and related features such as Tense. Pargram takes advantage of these cross-linguistic facts: imposing parallelism requirements and requiring different analyses to be empirically motivated on the basis of syntactic differences

Pargram is built on the XLE (Xerox Linguistic Environment) platform, developed and maintained at PARC, which implements LFG theory. It outputs *c-structures* (trees) and *f-structures* as the syntactic analysis. Its ambiguity management system takes advantage of its output in a packed representation of all possible solutions, which allows an application-specific disambiguator to avoid unnecessarily enumerating all solutions, while preserving the ambiguity.

On the basis of Pargram development so far, it is estimated that two person years are needed to create a grammar within Pargram with good coverage and depth (Kim et al. 2003); that is, one that can process sentences of realistic length and complexity. By joining Pargram, our project will benefit from its stable mature, well-supported grammar, and development environment, and from feedback from partner grammar developers in the Pargram research group. It reduces the need for software development. Instead, we will be able to focus on grammar research and development.

3 I-Pargram and corpus development

The I-Pargram project consists of three intertwined sub-projects: corpus development, lexicon development, and grammar development. In what follows, we describe the aims of I-Pargram, the need for Indonesian corpus development, the corpus architecture we envisage, and possible challenges of the envisaged corpus.

3.1 Aims of I-ParGram

Our research in I-ParGram has the following descriptive, typological, theoretical and computational linguistic aims.

- i. *Describing primary data*: to produce empirically verifiable corpus-based descriptions of meaning and structure that cover a broad range of constructions in different registers and genre variations;
- ii. *Building a balanced corpus* (see §3.2 and §3.4 below);
- iii. *Explicating precise linguistic information in the lexicon and grammar*: to provide explicit representations of grammar in the standard LFG representation (Bresnan 1982, 2001) and in the LFG semantics (Dalrymple 2001; Andrews 2003, in press) for computational processing, including the richness of lexical meanings in the corpus (Fillmore & Atkins 1994);
- iv. *Richer linguistic analyses*: (iii) is expected to lead to stronger and deeper linguistic-typological analyses of Indonesian data;
- v. *Improved linguistic analyses by rigorous testing*: to work out computational grammars containing rules in Pargram, featuring broad coverage of structures, efficient processing and high quality output, and representing the analyses in (iv).

3.2 The significance of corpus development

Achieving aims (i), (iii)-(v) outlined in 3.1 requires using existing reference grammars and dictionaries of the language, and creating an electronic corpus to ensure adequate coverage and develop testbeds (Manning and Schütze 1999). We envisage a modern balanced electronic Indonesian corpus that has certain corpus architecture satisfying certain types of requirements essential both for our project goals and for long-term benefit of wider communities (see further below). The corpus will be publicly available, probably through the digital archive of Pacific languages, PARADISEC.

While there are some corpora available, they do not seem to meet the specifications that we want to have. Therefore, it is significant that we develop the required corpus.

Generally speaking, the existing Indonesian corpora are not ‘balanced’ in terms of representative genres and register variation, most are not publicly accessible, and none comes with the required mark-up and annotation/tagging that we need. For example, a TREC-like corpus for Indonesian¹ from the online newspaper *Kompas* (Asian 2004) is restricted to written news only, and the MPI databases Jakarta Field Station² provide sample texts of child language acquisition corpus. Katakū (McKay 2004) is commercial and its underlying corpus/database is not available. There are also private text collections from individual linguists (Ewing 2003).

Identifying gaps in coverage of genre is an important step in our corpus development project. This will allow us to develop a framework for a corpus balanced in terms of representative genres and, if possible, register variation, and to set realistic limits to the range of language we expect the Pargram to deal with. With the help of a bridging grant, we have started work assessing the gaps in coverage of Indonesian corpora

¹ <http://goanna.cs.rmit.edu.au/~jelita/corpus.html>

² <http://lingweb.eva.mpg.de/jakarta/database.php>

3.3 *Source materials*

The materials come from the acquisition of the materials from the existing corpora and our own data collection.

We are acquiring Indonesian corpora, including 72 hours of recording of Jakartan Indonesian from the University of Wellington. The recording consists of natural language use of colloquial (Jakartan) Indonesian in a range of settings. We will negotiate with other corpus holders for further acquisition.

In addition, we will use publicly available electronic materials on Indonesian (including the internet), to create subcorpus consisting of balanced samples representing different text types of written Indonesian. Written Indonesian broadly reflects standard/formal Indonesian. However, we are also interested in the range of the variation of the constructions in spoken Indonesian, commonly regarded as non-standard Indonesian (Gil 2003).

At this stage of the project, we will focus on spoken Jakartan Indonesian because there are existing corpora, and because it is a well-known variety which is used in the media.

3.4 *Corpus design requirements*

Our large-scale Indonesian corpus must satisfy the following requirements (cf., Dipper et al. 2004): i) requirements due to the (linguistic) data content, ii) requirements pertaining to annotation and metadata, iii) requirements due to different users, and iv) technical requirements for long-term use. These requirements call for a well-designed corpus that is flexible for incremental addition of a range of text materials on the one hand, but is also maximal in conforming standards in digitization and annotation on the other hand.

3.4.1 *Data and corpus composition*

The following requirements are due to the nature of the (linguistic) data content:

Balanced and representative. The corpus is a large repository of different kinds of texts representing language use in a range of contexts with relevant parameters including (sub)genres/register, sex and age group of speakers, modalities (spoken vs. written), social contexts, event structures (monologue/dialogue), channel (radio broadcasting, TV, telephones, ...) etc.

Variation in size across text types is unavoidable, especially at the initial stages of corpus development.

Multi-media resources. Certain texts may come with audio and video files which need time alignment annotation.

Multi-lingual alignment. Formal Indonesian and Colloquial Jakartan Indonesian may differ significantly in certain aspects of their lexicon and grammar. In addition, there may be parallel texts in formal Indonesian and English, e.g. manual texts. Alignment of units of texts across languages is useful for research purposes and application in machine translation.

Linking to external resources. Certain acquired texts in the corpus may be linked to external resources.

3.4.2 Metadata and annotation

Metadata, meta-annotation. This is the information about the data and the annotation. Our corpus will include as complete metadata as possible, including the description of the content (topic, genre, communicative contexts, modalities, etc.), the people involved and their roles (collector of the data, the language speakers/consultants, the annotator, tools used, etc.), the associated media files. We will follow the OLAC (Open Language Archives Community) metadata set standard.³

Linguistic annotation. This can be of different kinds depending on research questions. For our purposes, we envisage linguistic annotation of the corpus in terms of grammatical (morphological and syntactic), lexical tagging and pragmatic tagging relevant for our LFG research, e.g. POS tagging and treebanking with grammatical and discourse functions.

Annotation depth. The extent of linguistic annotation depends very much on research questions. While the richer the annotation the better, it is very time consuming if done manually. We plan to have (semi-)automatic annotation for certain annotation process, but we anticipate quite a lot of manual tagging at the initial stage of the project.

We envisage a corpus that contains three sub-corpora with different degrees of annotation depth: i) sub-corpus 1, minimally annotated with some basic metadata annotation, ii) sub-corpus 2 with basic interlinear annotation in addition to the metadata, and iii) sub-corpus 3 with rich grammatical, lexical and pragmatic annotation for the purpose of our project, in addition to metadata and basic interlinear information. The three sub-corpora will increase in size over time throughout the project (Figure 1) with sub-corpus 1 will be likely to continue to grow.

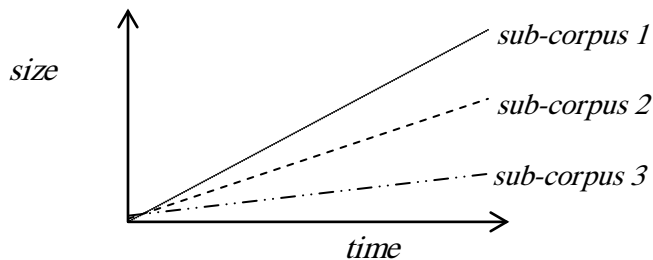


Figure 1: composition of the planned Indonesian Corpus

3.4.3 Users: presentation requirements

While we have our own research goals in building the Indonesian Corpus, the corpus itself is designed to be useful for scholars in different fields, not only syntacticians but also lexicographers, sociolinguists, and other interested parties. The following should be taken into considerations for the presentation of the corpus to meet wider public needs.

Web presentation. The ideal visual web-interface presentation should allow different options of viewing, e.g. text only and relevant portion of the text with certain annotation. Additional external viewers such as pdf must be also supported.

³ <http://www.language-archives.org/OLAC/metadata.html>

Search facilities. Search facilities should allow different (simple and complex) queries, including search for a word/word form, collocation, certain annotation or full text.

Export. The search results and corpus presentation must have export options which allow the user to further process and manipulate the data for his/her purpose.

3.4.4 Long-term use: technical requirements

The corpus will comply with certain linguistic and IT standards to ensure that the corpus can be accessed and further developed in future, to ease the application of external tools, and to allow exchange with other corpora.

4 Final remarks

4.1 Challenges

Given the goals, nature and requirements of the Indonesian corpus we want to develop, there are at least three main challenges to face in the course of the project.

One is how to know and evaluate that the large corpus is indeed balanced and representative. The difficult task would be to know that we have the right sampling to ensure representativeness and balance in the corpus. This requires good knowledge and understanding of the full extent of uses or genres of Indonesian and Jakartan Indonesian by which we can judge the representativeness of our corpus. Unfortunately, we do not yet (and perhaps will never) know this. Furthermore, to have the correct balance, we need to know frequency and importance of text varieties. This is also an issue that is not easy to determine in an objective way. We therefore interpret balance and representativeness in relative terms (McEnery, Xiao, and Tono 2006).

Another (technical) challenge is to design and implement a corpus architecture that meets our current needs/requirements with respect to certain features and standards as described in §3.4.3-3.4.4, but is also flexible enough for addition in response to different uses and standards in the future. The problem is that web-based technology and external tools are changing fast. We consider a web-based design that meets open-source standards (e.g. data structures with XML compliance) is the right way to go.

Designing a morphological analyzer, and also providing grammatical annotation, for our corpus (i.e. for the richly annotated sub-corpus 3) is also a challenge. Ideally we have a single morphological analyzer for both formal Indonesian and Colloquial Jakartan Indonesian. However, Sneddon (2006:1) mentions that Colloquial Jakartan Indonesian is ‘in many ways significantly different from the formal language ... to the extent that it deserves separate consideration’. An important feature of Colloquial Jakartan Indonesian is its verbal morphology (e.g. *-in and N-*) and a range of discourse (clitic) particles (e.g. *=kok, =kek*) which are absent in formal Indonesian. Furthermore, certain constructions in Jakartan Indonesian could be regarded unacceptable in formal Indonesian. The challenge would be then how to capture these register and acceptability variations precisely in our annotation. This is important if we want to design a unified system for (semi-)automatic annotation which can detect the variations. Given our goal to develop a large-scale corpus, we certainly want to have (semi-)automatic annotation to speed up sub-corpus 3 development.

4.2 Collaborative research

In this paper we have outlined our corpus development project and possible challenges ahead. The corpus development is part of wider project in the implementation of ParGram to Indonesian data. Our I-Pargram involves collaboration of experts in linguistics, grammar engineering, and computer science.

Our project brings together a multi-site and interdisciplinary team of investigators; each has relevant expertise to contribute to ensure the success of the project: Jane Simpson (Linguistics, Sydney University, Australia), I Wayan Arka and Avery Andrews (Linguistics, ANU, Australia), Mary Dalrymple (Linguistics and Philology, Oxford University, U.K/PARC, California), and Indonesian researchers (Maruli Manurung and Mirna Adriani) from The Faculty of Computer Science Universitas Indonesia Jakarta.

In terms of the **grammatical framework**, we will use our experience in developing grammars within the Lexical Functional Grammar (LFG) framework to create and test the grammar rules. In terms of the **computational linguistic side** of the program, Andrews and Dalrymple have long experience of connecting LFG grammars and computational approaches. Dalrymple was previously the manager of the whole Pargram project, and will train the others in XLE. Maruli and his colleagues in Jakarta together with the computational linguistics research associate and the data manager/programmer (to be recruited) will carry out the work of developing morphological analyzer and implementing a PARGRAM. In terms of the **languages**, the main responsibility belongs to I Wayan Arka, who is a native speaker of Indonesian. He will work with other native speakers of Indonesian from the research team in Jakarta. In terms of the **corpus development**, PI Dalrymple has experience adapting corpora to computational use, and CI Simpson has experience managing electronic corpora from the Aboriginal Studies Electronic Data Archive.

References

- Andrews, Avery. 2003. Glue logic vs. spreading architecture in LFG. *Proceedings of the ALS conference 2003*.
- . in press. Input and Glue in OT-LFG. In *Architectures, rules, and preferences: A festschrift for Joan Bresnan*. , edited by J. Grimshaw, J. Maling, C. Manning, J. Simpson and A. Zaenen. Stanford CA: CSLI.
- Asian, Jelita, Williams, Hugh E. and Tahaghoghi, S.M.M. . 2004. A testbed for evaluating Indonesian text retrieval *Proceedings of the 9th Australasian Document Computing Symposium, Melbourne, Australia, 13 December 2004*. <http://goanna.cs.rmit.edu.au/~jelita/corpus.html>.
- Bresnan, Joan. 2001. *Lexical functional syntax*. London: Blackwell.
- , ed. 1982. *The mental representation of grammatical relations*. Cambridge, Massachusetts: the MIT Press.
- Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar Project. *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation:1-7*.
- Dalrymple, Mary. 2001. *Lexical Functional Grammar, Syntax and semantics*. San Diego: Academic Press.

- Dipper, Stefanie , Lukas Faulstich, Ulf Leser, and Anke Lüdeling. 2004. Challenges in Modelling a Richly Annotated Diachronic Corpus of German. Paper read at Workshop on XML-based richly annotated corpora, LISBON, Portugal, May 2004.
- Ewing, Michael C. 2003. Affix and Affect in Indonesian *ISMIL 7: The Seventh International Symposium on Malay/Indonesian Linguistics Nijmegen, The Netherlands, 27-29 June 2003*
<http://email.eva.mpg.de/~gil/ismil/7/abstracts/ewing.html>
- Frank, Annette. 1999. From parallel grammar development towards machine translation. Paper read at Proceedings of MT Summit VII.
- Gil, David. 2003. Colloquial Indonesian dialects: How real are they? . *ISMIL 7: The Seventh International Symposium on Malay/Indonesian Linguistics Nijmegen, The Netherlands, 27-29 June 2003*
<http://email.eva.mpg.de/~gil/ismil/7/abstracts/gil.html>.
- Kaplan, R.M., and J. Bresnan. 1982. Lexical-Functional Grammar: A Formal System of Grammatical Representation. In *The mental Representation of Grammatical Relations*, edited by J. Bresnan. Cambridge, Massachusetts: the MIT Press.
- Kim, Roger, Mary Dalrymple, Ronald M. Kaplan, and Tracy Holloway King. 2003. Porting gramars between typologically similar languages: Japanese to Korean. *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation (PACLIC 17)*.
- Manning, Christopher, and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.
- McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-based language studies*. London: Routledge.
- McKay, Helen. 2004. Experiences in marketing MT for Indonesian. . *The Guide from Multilingual computing and technology: Going global: Asia 61 supplement*.
- Sneddon, James. 2006. *Colloquial Jakartan Indonesian*. Canberra: Pacific Linguistics.