

Challenges of developing a balanced and representative corpus for Indonesian ParGram

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

<http://rspas.anu.edu.au/linguistics/projects/iwa/IndProject/Web-IndonesianProject.htm>

Outline of the talk

- Introduction
 - Aims of the talk
 - Our collaborative research group
- ParGram & I-ParGram
- Corpus development
 - Its significance for our research
 - Corpus design & requirements
 - Expected challenges
- Plan and implementation?

Introduction

- Aims of the talk
 - To present our proposed project
 - building a balanced & representative electronic corpus of Indonesian
 - 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):
 - I-ParGram (Indonesian Parallel Grammar)
 - To initiate & foster
 - (future) research collaboration, particularly from relevant Indonesian institutions
 - To share experiences, expertise & resources
 - Contribution/acquisition of corpus materials

I-ParGram: collaborative research

- a multi-site and interdisciplinary team of international investigators

Personnel	Ling/LFG	Comp. Ling	Indo lgs	Notes
Jane Simpson	√			Linguistics, Sydney U, Australia
Wayan Arka	√		√	Linguistics, RSPAS ANU, Australia
Avery Andrews	√	√		Linguistics, CASS ANU, Australia
Mary Dalrymple	√	√		Linguistics, Oxford U, UK
Ruli Manurung, Bobby Nazief, Mirna Adriani		√	√	Computer Sc. UI, Indonesia
NLTT, PARC	√	√		Palo Alto, USA

The ParGram Project

<http://www2.parc.com/istl/groups/nltt/Pargram/>

- an international collaborative research project for the development of large-scale computationally tractable grammars and lexicons of the world's (major) languages
 - The underlying linguistic framework: LFG
 - Kaplan & Bresnan (1982), Bresnan (2001), Dalrymple (2001)
 - Built on the XLE (Xerox Linguistic Environment) platform:
 - Developed & maintained by NLTT, PARC
 - Implements LFG theory
 - Develop and process grammars in parallel
 - Current languages in the project: English, German, French, Japanese, Norwegian, Urdu, Welsh and Malagasy
 - Members include the corporate research laboratories & universities:
 - the Palo Alto Research Center (PARC) (USA), Fuji Xerox (Japan), and Stanford University, Oxford University, Manchester University, the Universities of Stuttgart and Konstanz, (Germany), the University of Bergen (Norway), the University of Essex (UK), and Langue et dialogue (France).

I-ParGram

- Joining ParGram:
 - 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, and
 - we can focus on grammar research and development
- Corpus-based computationally tractable lexicon and grammar with rigorous analysis and testing: two related expected results:
 - an empirically robust and linguistically well-grounded analysis, more in depth analysis of the grammar and lexicon than is currently available
 - a robust and sophisticated electronic data base (richly annotated texts and lexicon) with grammatical parser
 - potential to serve as the basic computational tool for an ongoing range of purposes such as machine translation, text mining, dialect identification, text generation, language research and language teaching/learning etc.

The significance of corpus development

- to ensure coverage and testbeds (Manning and Schütze 1999) for our descriptive, typological, theoretical and computational linguistic aims in I-ParGram:
 - *Describing primary data*
 - *Explicating precise linguistic information in the lexicon and grammar (within LFG)*
 - *Richer linguistic-typological analyses*
 - *Improved linguistic analyses by rigorous testing: computational Indonesian grammar in Pargram*
- The available Indonesian corpora do not seem to meet the specifications that we want
 - Develop our own corpus
 - Acquire materials from the existing corpora
 - Assess and fill in the gaps (with our own data collection)

Corpus design requirements

- Requirements due to the (linguistic) data content and composition
- Requirements pertaining annotation and metadata
- Requirements due to different users
- Technical requirements for long-term use

Corpus design requirements (cont)

- i) requirements due to the (linguistic) data content and composition
 - *Balanced & representative*
 - *Multi-media resources*
 - *Multi-lingual alignment*
 - *Linking to external resources.*
- ii) requirements pertaining annotation and metadata
 - *Metadata, meta-annotation*
 - *Linguistic annotation*
 - *Annotation depth*: three sub-corpora with different degrees of annotation depth

Corpus design requirements (cont)

- iii) requirements due to different users
 - *(Interactive) Web presentation*
 - *Search facilities.*
 - *Export*
- iv) technical requirements for long-term use
 - comply with certain linguistic and IT standards
 - accessed and further developed in future
 - to ease the application of external tools, and
 - to allow exchange with other corpora

Challenges

- How to know and evaluate that the large corpus is indeed balanced and representative
 - Requires good knowledge & understanding of the full extent of lang uses: (sub)categories, frequency and importance
 - In relative terms
 - Related issues of text acquisition: availability, legal (copy right) matters, time & funding constraints
- How to design and implement a corpus architecture that meets our current needs/requirements with respect to certain features and standards, but is also flexible enough for addition in response to different uses and standards in the future
 - Open-source standards
- Designing a morphological analyzer for Standard Indonesian and Colloquial (Jakartan) Indonesian, and also providing rich annotation (sub-corpus 3)
 - A single morphological analyzer for both?
 - How to speed up rich (and good) annotation? manual vs. semi-automatic
 - Time and funding constraints

The plan ...

- Funding:
 - Sydney U grant
 - ARC: 2008-2011?
 - Possible Linkage grants?
- Recruitment, further collaboration & implementation
 - Visit to Jakarta
 - Fakultas Ilmu Komputer UI
 - Pusat Bahasa
 - LIPI
 - ...
 - Individuals?
 - From ISMIL participants?
 - Prospective PhD students?
 - Prospective data manager & Indonesian team members
- Project homepage:
 - <http://rspas.anu.edu.au/linguistics/projects/iwa/IndProject/Web-IndonesianProject.htm>

Composition of the planned Indonesian corpus: three sub-corpora

