THE ELEVENTH
INTERNATIONAL SYMPOSIUM ON MALAY/INDONESIAN LINGUISTICS

6-8 August 2007
Manokwari, Indonesia

Challenges of developing a balanced Indonesian corpus

I Wayan Arka (ANU)
Wayan.arka@anu.edu.au

Jane Simpson (The University of Sydney)
jhs@mail.usyd.edu.au

Abstract

We argue for the importance of having a balanced corpus, corpus-based descriptions/analyses of meanings and structures, and precise explication of linguistic information for the purpose of computational application of the descriptions/analyses. We will outline our proposed project of building a balanced 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 will discuss the nature of the corpus that we will develop, and the types of coding required for different analyses. We will assess existing corpora, outline the research plan and discuss challenges that we will face.

Parallel Grammar Project: http://www2.parc.com/istl/groups/nltt/pargram/