This activity aims to bring together linguists, academicians, computer scientists, publishers, language educators, and government agencies willing to share their knowledge and work together towards building the Philippine corpus. At the end of the workshop, the participants will have identified the target initial contents of the corpus, discussed the possible methodologies for this undertaking, and explored avenues for cooperation to the Philippine corpus project.

The long-term goal is the establishment of a lasting, representative and multi-purpose record of all languages spoken in the Philippines, which shall include not only written texts but also audio and video recordings transcribed, annotated, compiled and digitized in a structured corpus made accessible to a broad group of users.

Natural Language Processing Academic Area
College of Computer Studies, DLSU-M

The College of Computer Studies, De La Salle University-Manila, created the Academic Area on Natural Language Processing (AA-NLP) to support and to promote research in the field. Research areas of the AA-NLP include NL understanding, NL generation, machine translation and information retrieval, among others. To know more about us and our activities, you may visit our website at http://ccs.dlsu.edu.ph/nlp-team

Komisyon sa Wikang Filipino

The Komisyon sa Wikang Filipino (Commission on the Filipino Language) is the official research agency for Philippine languages. It was established in accord with the 1987 Constitution of the Philippines and earlier in the 1930s as the Surian ng Wikang Pambansa. The KWF seeks to become the center of scholarship in Philippine languages and literatures. Its mission is to develop, promote and preserve the languages used by Filipinos in various domains. Its mandate is to undertake, coordinate and support linguistic and literary studies in the national, local or foreign languages.

For inquiries, please contact the KWF through
Dr. Ricardo Ma. Nolasco, at rnolasco_upmin@yahoo.com or
Dr. Sheilee Boras-Vega, at shee_bvega@yahoo.com or
at (02) 7362521 or (02) 7363832

or DLSU-M through
Dr. Rachel Roxas, at roxasr@dlsu.edu.ph or
Ms. Glennifer Ramos, at fhey_0325@yahoo.com or
at (02) 536-0278

De La Salle University – Manila
Advanced Research Institute for Informatics, Computing and Networking

in cooperation with

Komisyon sa Wikang Filipino

presents

A Consultative Workshop (Part II):
Towards Building the Philippine Corpus

Yuchengco Hall 408
De La Salle University-Manila
2401 Taft Avenue
1004 Malate, Manila

8:30 am to 3:00 pm, November 28, 2007
Towards Building the Philippine Corpus
by Dr. Rachel Roxas
A Jonathan Dy Professorial Chair Lecture

Abstract
The vision of building the Philippine corpus will be presented. A
Corpus is a collection of documents in a particular language or
languages that are to be stored, managed and analyzed in digital
form. It is indispensable in researches on various aspects of the
study of the nature and functioning of natural language, and its multi-
faceted applications such as language education, lexicography, and
natural language processing. Many corpora have been built for
major languages of the world and for some Asian minor languages,
but there is quite limited work on building one for our Philippine
languages. The information in a corpus can be wide-ranged, and
will transcend time and space, and will be across languages,
regions, genres and forms. Both historical and contemporary
information can be stored in the corpus. All Philippine languages
both major, minor and near extinction are to be documented; thus,
the Philippine corpus will be across Philippine regions. It will
consider different genres (newspaper articles, literary works,
folklore, etc) and different forms (text, speech and video). It aims to
transcend physical barriers, enabling contribution to the corpus and
corresponding access to this information across land and sea
through the internet. For the fulfillment of this vision, a concerted
effort is necessary to populate, annotate and verify the corpus.