Cooperation on Language Resources Sharing – Perspective from Asian languages –

Tokunaga Takenobu
Tokyo Institute of Technology

Past (My experience)

- RWC text database (1994-2001) (LREC 1998)
  - News paper articles (5 years)
    - morphological/syntactic/semantic/discourse tags (GDA)
    - category code (UDC)
  - *Iwanami* Japanese dictionary (60,000 words)
    - morphological/syntactic tags

Present

- Survey of Asian linguistic resources
  - Following OLAC framework
  - under the frame work of Asia Federation of Natural Language Processing (AFNLP)
    URL: http://www.aisanlp.org
- Prototype of resource DB → Screen shot
  URL: http://tokunaga-www.cs.titech.ac.jp/ALR/search.html
  - Japanese 56 entries
  - Korean 10 entries

Example of Japanese Resource

- Thesaurus
  - Bunrui Goi Hyo (NIJL) 32,600
  - NTT thesaurus 300,000
- Dictionary
  - IPAL lexicons/ Verb:861, Adjective:136, Noun:1,081
  - EDR dictionaries/ Word (260,000), Bilingual (JE:240,000, EJ:160,000), Concept (410,000), Co-occurrence (930,000), Technical terminology (197,000)
- Tagged corpus
  - RWC DB
  - Kyoto-U corpus 40,000 sents.
  - EDR corpus 200,000 sents.

Future

- Interrelations between existing resources
  - Resources of same language (eg. linking different kind of resources)
  - Resources of cross language (eg. the EAGLES proposal on Asian language)
- Critical mass to be a “real” standard
- Demand on minority language resources?