Toward an Ontology-based Language Service Infrastructure

Yoshihiko Hayashi
Osaka University and NICT
mailto:hayashi@lang.osaka-u.ac.jp
Outline

- Language Infrastructures
- The Language Grid project
- Language Service Ontology
- Summary and Discussions
Language Infrastructures

- **Research Infrastructures**
  - **CLARIN** is committed to establish an integrated and interoperable research infrastructure of language resources and its technology. It aims at lifting the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and therefore enabling e-Humanities

- **Service Infrastructures**
  - **Language Grid** provides a language infrastructure on which language services that are useful in intercultural collaboration can be composed, delivered, and utilized
Language Grid

- Connecting World's Language Services to Support Intercultural Collaboration
- a project at NICT (National Institute of Information and Communications Technology)
Remark on LS and LR

- LS (Language Service): a Web service whose functionality is somehow related to language
- LR (Language Resource) in a broader sense can be classified into:
  - static data resource
    - corpus, lexicon, language model, etc.
    - *but, accessed via a processing resource (accessor)*
  - language processing resource
    - NLP tool/system
    - Language resource accessor
“Layer cake” of the Language Grid

Intercultural Collaboration Tools

Language Services (as composite Web services)

Language Resources (as atomic Web services)

P2P Grid Infrastructure

Set of standardized APIs are required.

Also, functional meta-descriptions for atomic services are necessary.
Web Services - current figure

- Composite Web Services (Workflows)
  - Domain Specific Translation
  - Multi-hop Translation
  - Back Translation
- Atomic Web Services
  - Morphological analyzers/POS taggers: Japanese, Chinese, Korean, English, German, Spanish, French, Italian, Dutch, Russian, Bulgarian
  - Bilingual dictionaries: Life science terms (Japanese, English), Disaster management terms (Japanese, Chinese, Korean, English, French, Spanish), Academic terms (Japanese, English)
very top of the Language Service Ontology

Language Process

Language Process (Service) Triangle

Language Process

Language Process (Service) Triangle

Top-level of the Language Service Ontology

Language Data

Linguistic Object

Language Process

Language Process (Service) Triangle

Language Process

Language Process (Service) Triangle

Top-level of the Language Service Ontology

uses

processes

contains-of

Language Process

Language Process (Service) Triangle

Language Process

Language Process (Service) Triangle

Top-level of the Language Service Ontology

uses

processes

contains-of

Language Process

Language Process (Service) Triangle

Language Process

Language Process (Service) Triangle

Top-level of the Language Service Ontology

uses

processes

contains-of

Language Process

Language Process (Service) Triangle

Language Process

Language Process (Service) Triangle

Top-level of the Language Service Ontology
Considerations in Developing the Sub-ontologies

- They shouldn’t be incompatible with the relevant standards
- **LAF** (Linguistic Annotation Framework) together with related standards: should be incorporated for defining the types of input/output data of processing resources
- **LMF** (Lexical Markup Framework): should be introduced for developing a taxonomy of lexicons, as well as linguistic information encoded in lexicon entries
“Ontologization” of LAF
SyntacticAnnotation as a subclass of LinguisticAnnotation
Considerations in Developing the Sub-ontologies

- They shouldn’t be incompatible with the relevant standards
- **LAF** (Linguistic Annotation Framework) together with related standards: should be incorporated for defining the types of input/output data of processing resources
- **LMF** (Lexical Markup Framework): should be introduced for developing a taxonomy of lexicons, as well as linguistic information encoded in lexicon entries
“Ontologization” of LMF (core model)
Sketch of the LMF-based Lexicon Ontology
Configuration of ConceptLexiconEntry
Summary

- Language Grid as a service-oriented language infrastructure
- Language Service Ontology as a shared ground for describing atomic Web services (wrapped language resources)
  - "Ontologization" of relevant ISO standards
- The ontology has been developed in cooperation with DFKI Language Technology (Germany) and ILC-CNR (Italy)
- ... the current set of Web APIs and the ontology is not directly linked
Discussions

- How to adjust/control abstraction level of the linguistic annotations?
  - handle idiosyncrasies; is theory-neutral a sufficient solution?
  - avoid unnecessary details, depending on the purpose of a language service

- Necessity of an ontology for linguistic attributes/values
  - relevant standards: Data Category (Registry)
    - DC should be more structured and the DCRs should be distributed over the Web
  - also required for ULA?
Thank you!

- Come visit us at: http://langrid.nict.go.jp/

- References:
IC Collaboration Tools

**Langgrid Chat**
A multilingual communication tool for chat.

- Users can read and write in their first language, so they can communicate easier.

**Case:** NPO Pangaea
Pangaea tries to create bond among kids around the world via online universal playground. Langgrid Chat is used by volunteers to plan activities in order to connect different countries.

**Langgrid Blackboard**
A multilingual blackboard for information sharing among people in different countries.

- Information on the cards can be shared in different languages.

**Case:** All For One Collaboration Project
This tool has been used at seminars in universities or research institutes, to help exchange students and foreign researchers who need language support.

**Langgrid Input**
A multilingual text input tool for existing tools like BBS.

- User can edit specialized dictionary for specialized translation.

**Case:** NGO JEAR
Children around the world used Langgrid Input to create English messages for the official BBS of the “Natural Disaster Youth Summit.”
Composite Service Workflow: multilingual specialized translation

Please look for the shelter in the neighborhood.

* “Shelter” is registered in a specialized dictionary

Multilingual Back Translation Workflow

En to Ja Specialized Translation Service
En to Fr Specialized Translation Service
En to Ko Specialized Translation Service
Ja to En Specialized Translation Service
Fr to En Specialized Translation Service
Ko to En Specialized Translation Service

En-Ko Specialized Translation Workflow

English Morphological Analysis Service

Any specialized term?

YES

Specialized Dictionary Service

NO

En to Ja Translation Service
Ja to Ko Translation Service
En to Ko Translation Service

En-Ko Translation Workflow

Ja: どうぞ、付近の避難所を捜しなさい。
Fr: Veuillez rechercher le refuge dans le voisinage.
Ko: 가까이 있는 피난소를 찾아 주십시오.

Translation

Back Translation

Please, look for the near shelter.
Please search for the shelter in the neighborhood.
Please look for a close shelter.

Atomic Services

Tree Tagger
(Universitat Stuttgart)

Multi-language Glossary on Natural Disasters
(Jp, En, Ko, Zh, Es, Fr)
(Ritsumeikan Univ.)

Web-Transer
(Cross Language, Inc.)

J-Server
(Kodensha Co., Ltd.)

Translation

Ja : Japanese
En : English
Ko : Korean
Fr : French
Core and Service Nodes in P2P Grid Architecture