Ontologies for a Global Language Infrastructure

Yoshihiko Hayashi
Graduate School of Language and Culture, Osaka University
Toyonaka, 560043 Osaka, Japan
hayashi@lang.osaka-u.ac.jp

Thierry Declerck
DFKI GmbH,
Language Technology Lab
D-66123 Saarbrücken, Germany
declerck@dfki.de

Paul Buitelaar
DFKI GmbH,
Language Technology Lab &
Competence Center Semantic Web
D-66123 Saarbrücken, Germany
paulb@dfki.de

Monica Monachini
Instituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1-56124 Pisa, Italy
monica.monachini@ilc.cnr.it

Abstract

Given a situation where human language technologies have been maturing considerably and a rapidly growing range of language data resources being now available, together with natural language processing (NLP) tools/systems, a strong need for a *global language infrastructure* (GLI) is becoming more and more evident, if one wants to ensure re-usability of the resources. A GLI is essentially an open and web-based software platform on which tailored *language services* can be efficiently composed, disseminated and consumed. An infrastructure of this sort is also expected to facilitate further development of language data resources and NLP functionalities. The aims of this paper are twofold: (1) to discuss necessity of ontologies for a GLI, and (2) to draw a high-level configuration of the ontologies, which are integrated into a comprehensive *language service ontology*. To these ends, this paper first explores dimensions of GLI, and then draws a triangular view of a language service, from which necessary ontologies are derived. This paper also examines relevant ongoing international standardization efforts such as LAF, MAF, SynAF, DCR and LMF, and discusses how these frameworks are incorporated into our comprehensive language service ontology. The paper concludes in stressing the need for an international collaboration on the development of a standardized language service ontology.

1 Introduction

With the recent developments of the Semantic Web and progresses of the associated methodologies and standards, demands for an open and distributed infrastructure for sharing language resources and technologies can be addressed now on a new basis (Buitelaar et al., 2003; Calzolari, 2006). In this paper, we call such an infrastructure a *global language infrastructure* (GLI) GLI should accommodate language resources and technologies world-wide. A GLI thus should inherently address multilingual and multicultural issues.

More precisely, a GLI is an open and web-based software platform to which resources can be easily plugged in, and on which tailored *language services* can be efficiently composed, disseminated and consumed. Here a language service simply means a web service whose functionalities are generally related to human language; it can range from simple dictionary access to more complicated linguistic analysis, as well as conversion of linguistic expressions such as translation or paraphrasing.

We can mention the following initiatives/projects as examples of an obvious effort towards such a language infrastructure:

- *CLARIN*[1] is committed to establish an integrated and interoperable research infrastructure of language resources and technology. It aims at addressing the current fragmentation by offering a stable, persistent, accessible and extendable infrastructure that will enable the development of "e-Humanities".
- *Language Grid*[2] provides a language infrastructure on which language services that are useful in intercultural collaboration can be composed, delivered, and utilized. On the Language Grid, existing language data resources, NLP tools/systems and newly created community-based resources can be efficiently and effectively combined (Ishida, 2006). In addition, the Language Grid presents an operation model to address complicated issues associated with intellectual property rights and contracts (Ishida et al., 2008).

These two initiatives share issues of interoperability and reusability of language data resources and NLP tools/systems, even though their primary objectives are totally different. This calls for an opportunity to work out a common strategy for these crucial issues.

With this background, this paper argues that a GLI should be ontology-based, and presents a high-level configuration of the ontologies, which are integrated into a comprehensive *language service ontology*. This paper also examines relevant ongoing international standards, and discusses how these frameworks can be ontologized and incorporated into the comprehensive language service ontology.

2 Dimensions of GLI

2.1 Objectives of GLI

Needs for a language infrastructure have originally emerged from research fields including NLP and a range of e-sciences, which require mining from textual resources. For example, Klein and Potter (2004) presented two use cases; one is a workbench for NLP researchers, and the other is a text-mining tool for e-science researchers who are not necessarily NLP experts.

More recently, CLARIN explicitly targets its users to communities of e-humanities, and tries to offer its services to:

- The different communities of linguists to optimize their models and tools to the benefit of all who are using language material,
- Humanities scholars in the broad sense to facilitate access to language resources and technology, and
- The society as a whole to enable lower thresholds to multicultural and multilingual content.

In contrast, the Language Grid has been launched for providing a language infrastructure for supporting verbal, particularly cross-language, communications that are observed in activities of intercultural collaboration. To achieve this goal, the Language Grid provides an environment in which existing NLP tools/systems and newly created community-based language data resources can be efficiently combined. A number of communication tools are publicized on the project web site.

Here we should remark that (1) the user of a GLI is not necessarily an NLP expert, and (2) not only language data resources but NLP tools/technologies and their useful combinations are involved in a GLI.

2.2 Types of users in GLI

Users, or participants, of a GLI can be classified into the following types:

- A language resource provider who disseminates a language resource or NLP functionality in the form of a language service by wrapping it as a web service,
- A language service composer who composes a composite web service by combining atomic language services, and
- A language service end user who simply consumes a language service.

From a language infrastructure perspective, it is of crucial importance to provide useful support for a language resource provider in creating the wrap-

[1] http://www.clarin.eu/

[2] http://langrid.nict.go.jp/

pers, and for language service composers in authoring composite language services. To these ends, a standardized framework for describing language data resources and NLP tools/systems is strongly required (Hayashi, 2007).

2.3 Technical ingredients of GLI

As implied from the discussions so far, technical ingredients of a GLI are: (1) NLP tools/systems ranging from dictionary access systems and linguistic analyzers to machine translation systems, and (2) language data resources, such as lexicons or corpora. In addition to these, a GLI has to be aware of abstract linguistic objects such as linguistic expression, linguistic annotation or even linguistic meaning, because these types of abstract objects comprise the data to/from NLP tools/systems, as well as content of language data resources.

3 Ontologies for a GLI

3.1 Necessity of a comprehensive ontology

In principle, most of the existing language data resources and NLP tools/systems have been created independently, resulting in a situation where data format, annotation scheme, access method and other features are all idiosyncratic. This obviously will be a burden for establishing a GLI which ensures interoperability and reusability of language data resources and NLP tools/systems. To address this issue, standardization is inevitable: standardized APIs are necessary for NLP tools/systems; standardized data semantics as well as data format are required for language data resources. In addition and importantly, these standards should be designed based on a comprehensive shared ontology which covers all possible elements of a GLI.

3.2 Triangular view of a language service

In order to facilitate the development of a comprehensive ontology, it should be divided into appropriate sub-ontologies, each covering a grouped set of elements. Figure 1 shows a triangular view of a language service. Note that a language service is provided by a *language process*, not solely by language data or linguistic objects. Therefore language processes should be placed at the vertex of the triangle. A language process, in general, processes a linguistic expression which may or may not be linguistically annotated. We denote abstract objects such as linguistic expression or linguistic annotation as *linguistic object*. Linguistic objects may comprise a *language data* resource such as a corpus or lexicon; hence it would be utilized by a language process. This triangular view of a language service gives us a foundation on which necessary sub-ontologies are developed.

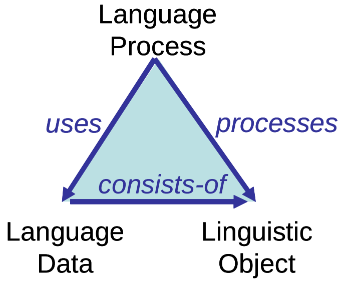

Figure 1: Triangular view of a language service.

3.3 Top-level of the language service ontology

Figure 2[3] illustrates the top-level of the language service ontology that is configured according to the language service triangle depicted in Fig 1. Each box in the figure denotes a top-level class in the ontology, which is defined in further detail by a sub-ontology. Among these top concepts, **`LanguageService`** is the top-most concept. As discussed with the language service triangle, a language service is provided by **`LanguageProcessingResource`** which takes **`LinguisticExpression`** as input/output and uses **`LanguageDataResource`**. Note that a language data resource does not provide a language service by itself; it is always used through an access mechanism which is an instance of some sub-class of the processing resource class.

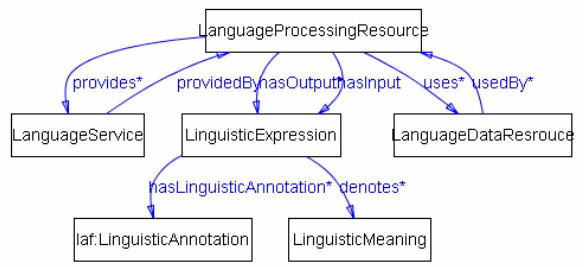

Figure 2: Top-level of the language service ontology.

[3] All the figures (except Fig.1) were produced with the OntoViz plugin of the Protégé ontology editor.

In further detailing the sub-ontologies, we believe it to be important to incorporate related international standards.In this sense, we have been looking at frameworks for linguistic annotation and lexicon modeling that have been discussed in international standardization bodies. The frameworks for linguistic annotation are incorporated into our ontology not only for specifying the input/output data of NLP tools, but also for defining the content of corpora. On the other hand, the framework for lexicon modeling is introduced to have a formal foundation for developing a taxonomy of lexicon classes, which are obviously subclasses of the language data resource class.

4 Ontology for Linguistic Annotations

Figure 2 also depicts an ontological configuration for abstract linguistic objects such as linguistic expression, linguistic meaning and linguistic annotation. It says: (1) a linguistic expression (**`LinguisticExpression`**) in a language denotes some meaning (**`LinguisticMeaning`**), even if it is not explicitly represented, (2) a linguistic expression should be the input or the output of a language process, and (3) a linguistic expression can be multiply annotated (**`LinguisticAnnotation`**). The last point is of crucial importance, because any framework for linguistic annotation has to be able to accommodate multiply layered annotations, given the possibility that the target linguistic expression would be annotated by more than one analyzer, each of which possibly doing its job on a different linguistic level. Among the linguistic objects, ontological configuration for the linguistic annotation should be most carefully designed with respect to the interoperability and reusability of language data resources and NLP tools, because the data to/from a linguistic analyzer, as well as the content of a language data resource should be represented as linguistic annotation.

Frameworks that are necessary for standardized linguistic annotations have been actively developed and disseminated by the ISO TC37/SC4[4] committee; these include LAF (Linguistic Annotation Framework) (Ide and Romary, 2006), MAF for morphosyntactic annotation (Clément and de la Clergerie, 2005), SynAF for syntactic annotation (Declerck, 2006), and others. Among these, the LAF is the most general *umbrella* framework, and the other frameworks inherit the basic properties of LAF. As these frameworks have not been defined in the form of an ontology, we decided to *ontologize* these frameworks and incorporate them into the language service ontology. Here to ontologize simply means to give OWL (Web Ontology Language) (McGuinness and Harmelen, 2004) specifications to relevant parts of the framework.

Figure 3 illustrates a high-level configuration of the sub-ontology for linguistic annotations. This configuration corresponds to the LAF framework. As shown in the figure, a linguistic annotation has a start position and an end position for designating the span of annotation in the target linguistic expression[5]. This allows us to implement so-called stand-off annotation, and hence enables multiple annotations on the same data set. It also accommodates a feature structure for representing the annotation content.

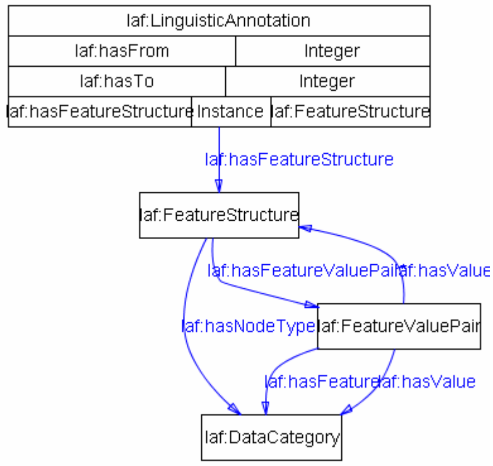

Figure 3: Configuration for LAF.

As noted in (Declerck et al., 2007), the LAF does not provide specifications for content categories; instead it includes a DCR (Data Category Registry) (Wright, 2004) that contains pre-defined data elements and schemas that may be used in annotations. Current configuration of the data categories does not induce taxonomical structure. Nevertheless the linguistic annotation class should be further organized into sub-classes based on which data categories should be included.

Figure 4 summarizes the ontological configuration for MAF and SynAF, introducing classes for

[4] http://www.tc37sc4.org/

[5] In LAF, this is called *primary data*.

segment (**SegmentAnnotation**), syntactic constituent (**SyntacticAnnotation**), and dependency relation (**DependencyRelation**). Note that these classes have been explicitly introduced, although these, in principle, should be represented with the feature structures. Although it is not depicted in the figure, the feature structure for representing morpho-syntactic annotation attached to a segment should be restricted to only include MAF conformant data categories. A similar story should apply to SynAF. As proposed in (Declerck, 2006), SynAF is designed to be able to represent two syntactic properties of a human language: *constituency* and *dependency*. Therefore the syntactic annotation class should be defined to have a specialized feature structure whose node type is restricted to the categories defined in the data category sub-profiles for constituency relation or dependency relation.

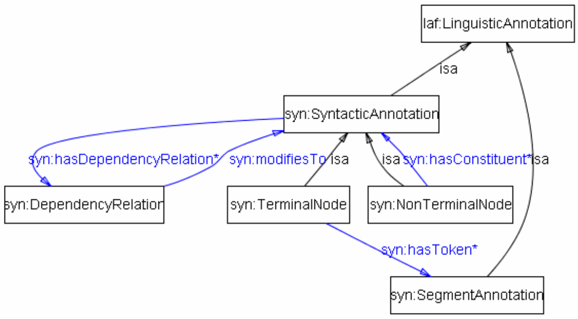

Figure 4: Configuration for MAF and SynAF.

With the ontology described so far, any linguistic expression in the proposing language service ontology can be typed according to the type of linguistic annotation it has. This type information can be effectively utilized in dynamic composition of composite services, in which checking of the input/output constraints given in the meta-description of a processing resource is necessary.

5 Ontology for Lexicons

The class for language data resource (**LanguageDataResource**) is currently organized by subclasses for corpus (**Corpus**) and lexicon (**Lexicon**). The corpus class can be further organized into subclasses according to the type of content, where type can be defined by the type of annotation of the content. Thus we can have an interrelation between the corpus ontology and the linguistic annotation ontology.

Similarly but not identically, the lexicon class should be organized into subclasses by the type of lexical content, and the type should be defined based on the features of a lexical entry in the target lexicon. Here we have an opportunity to incorporate ongoing standardization work in lexicon modeling into our language service ontology. To do this, we have first ontologized parts of the LMF (Lexical Markup Framework) (Francopoulo et al., 2006) which is also in the process of standardization by ISO TC37/SC4, and then connected these with the lexicon class taxonomy.

The ultimate goal of LMF, as stated in (ISO DIS 24613:2007), is to create a modular structure that will facilitate true content interoperability across all aspects of electronic lexical resources. The modular structure of LMF consists of a core package and a number of extensions for modeling a range of machine readable dictionaries (MRDs), and NLP lexicons. These LMF extensions are expressed by extending the LMF core package, encouraging us to ontologize them by organizing the classes defined in the core package as subclasses of the top LMF class.

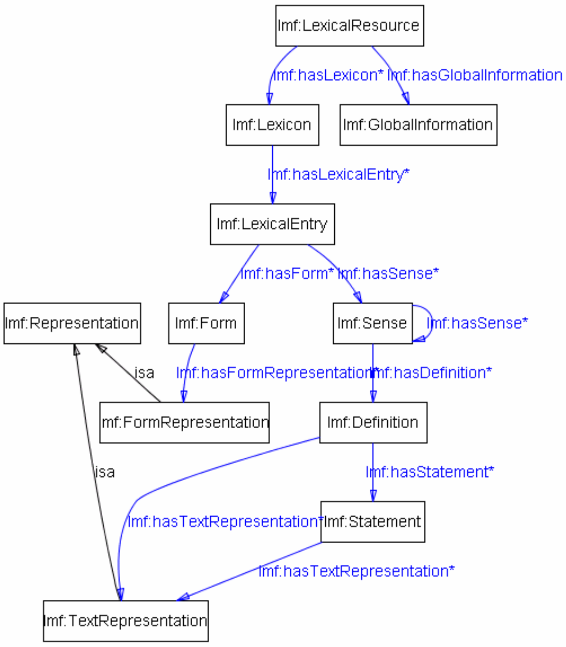

Figure 5: Configuration for LMF core model.

Figure 5 illustrates the ontological configuration for the LMF core model, while Figure 6 shows a part of the LMF NLP Semantics extension, which is associated in particular with the lexical semantic notions of the extension. As seen in these examples, ontologization of LMF is rather straightforward.

Then the questions should be:

- How can we define the taxonomy of the lexicon while referring to the ontologized LMF?
- How can we define a class for lexicon access, which is a sub-class of the processing resource (**`LexiconAccessor`**), while referring to the ontologized LMF?

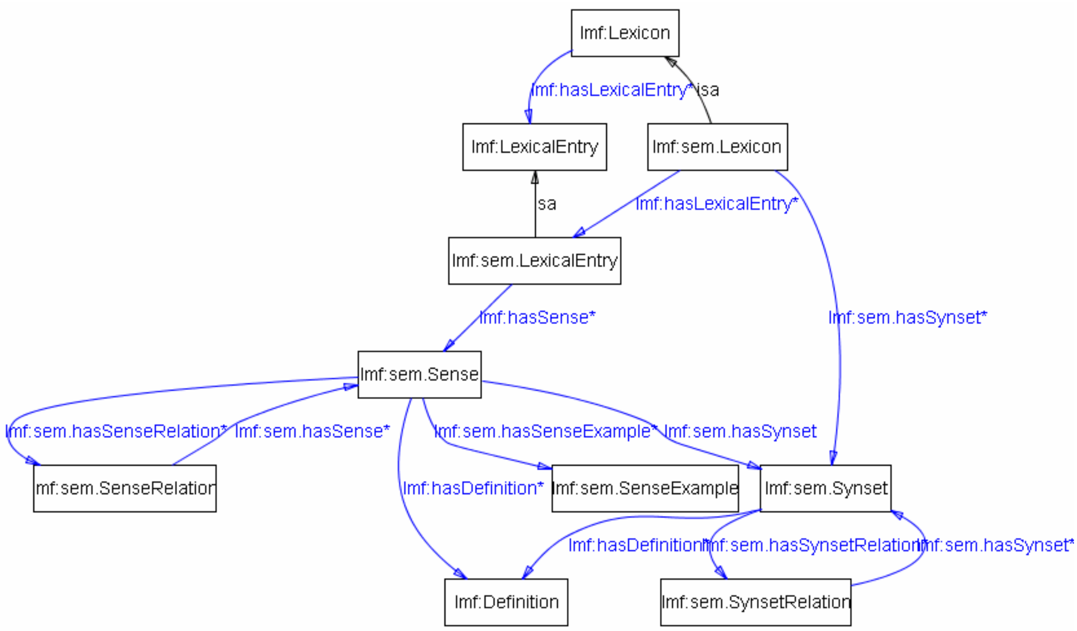

Figure 6: Configuration for LMF NLP Semantics.

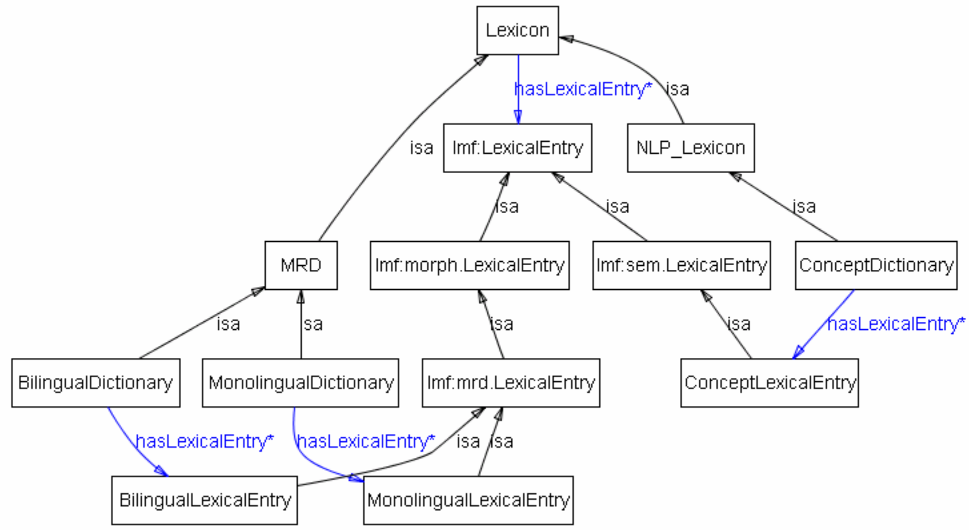

Figure 7: Taxonomy of lexicon class.

Figure 7 shows a taxonomy of the lexicon class, stating that each of the lexicon subclasses is defined in terms of the type of the lexical entries defined in the ontologized LMF. For example, **`BilingualDicctionary`**, a sub-class of **`MRD`**, is defined by **`hasLexicalEntry`** property whose range is restricted to **`BilingualLexicalEntry`**, which, in turn, is a descendant class of **`LexicalEntry`**. In order to incorporate some new type of lexicon, we have to first introduce a new sub-class, then appropriately place it some-where in the lexical entry class in the ontologized LMF and finally relate it to the lexicon taxonomy.

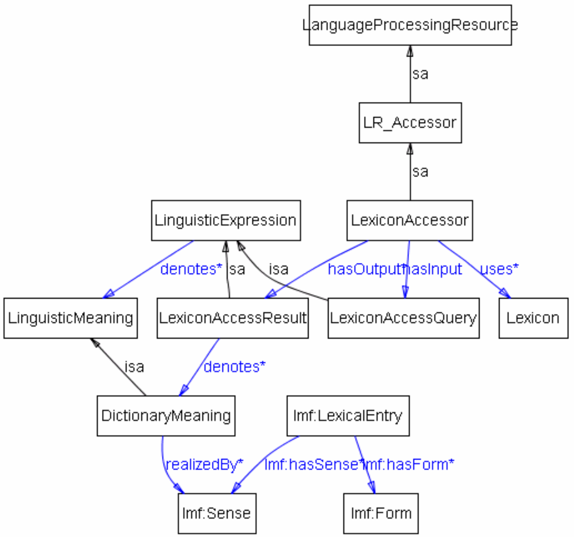

Figure 8: Configuration of lexicon accessor class.

Figure 8 summarizes the ontological definition for the lexicon accessor class; its input is restricted to a sub-class of the linguistic expression class (**LexiconAccessQuery**), whereas the output is restricted to **LexiconAccessResult** which is also a sub-class of the linguistic expression class. The former is defined to have properties for query conditions, while the latter is restricted to denote an instance of **DictionaryMeaning**, which is a sub-class of **Meaning**. Note that the dictionary meaning would be **realizedBy** an instance of the **Sense** class in the ontologized LMF. Here we have an explicit interrelation between the part of language service ontology with the LMF ontology. Note also that the **Sense** instance is associated with an instance of **LexicalEntry** class, and the associated **Form** instance should match with the linguistic expression given in the input query to the lexicon accessor. A *deep* constraint like this, however, is unfortunately beyond the representivity of the OWL formalism, hence not explicitly encoded. To encode such a deep constraint, the notion of *process* have to be introduced with a framework (e.g. SWRL) (Horrocks, et al., 2003) for expressing complicated logical relationships.

6 Related Work

Klein and Potter (2004) sketch an ontology for NLP services with OWL-S specifications. Their proposal unfortunately did not include ontologies for abstract linguistic objects such as linguistic annotations. Hayashi (2007) proposed a linguistic service ontology in the context of the Language Grid. Although it discussed a taxonomy for NLP tools, it did not present any details on the linguistic annotation and lexicon modeling.

LT World (Jörg and Uszkoreit, 2005) is a comprehensive knowledge portal for language technologies. One of the unique features of LT World is that it is based on a multi-dimensional ontology. For example, it classifies language technologies into such dimensions as: application, linguality, languages, technologies, linguistic area, and linguistic approach. This part of the ontology could be incorporated into our ontology especially for specifying the language processing resources.

Several relevant frameworks around language data resources have been actively developed by ISO TC37/SC4. As noted in this paper, we will carefully observe the activities, and incorporate the results as much as possible into our language service ontology. Among these, future development of the DCR will be of importance. That is, by developing an ontology for linguistic categories on top of the basic DCR data categories, we will have an opportunity to explicitly define relations among the data categories in our language service ontology. In this regard, our approach to the ontology for linguistic categories is in some degree different from the one taken by GOLD (Farrar and Langendoen, 2003), where not only linguistic categories but complex relations among them are fundamentally defined within the central ontology.

7 Concluding Remarks

A *global language infrastructure* (GLI) an open and web-based software platform on which tailored language services can be efficiently composed, disseminated and consumed. Given the increasingly realistic scenario in which language data resources and NLP tools/systems will be ubiquitous on the web, a comprehensive ontology (*language service ontology*) for describing these elements will be vital in addressing such issues in interoperability and reusability.

In this paper, we have examined a triangular view of a language service, which consists of language processing, language data, and linguistic objects. Based on this definition, we have presented a top-level ontology configuration along with an essential set of sub-ontologies; these include ontologies for processing resources, language data resources, linguistic annotations, and lexicons. Among these, the ontologies for linguistic annotations and lexicons have been substantially detailed while referring to the ISO frameworks LAF, MAF, SynAF, DCR, and LMF. In doing so, we ontologized an essential part of these frameworks, and incorporated them into our comprehensive language service ontology.

We strongly believe that although the results presented in this paper are still preliminary, the resulting language service ontology will be essential in defining an ontology-based GLI. Obviously, we still have to provide further detail for the presented sub-ontologies by looking at concrete language data resources and NLP tools/systems for a range of human languages. In parallel, we will need to develop an approach for handling any differences in desired expressiveness inherent to the objective of a GLI; e.g., a language research infrastructure may require precise linguistic descriptions, while an infrastructure for NLP applications might demand more coarse-grained linguistic descriptions, while focusing rather on detailed communicative aspects.

To conclude, in reaching an ontology-based GLI, we will need to establish a community of experts from a range of relevant research areas and human languages. We sincerely hope that this paper will contribute to the initiate such an initiative.

Acknowledgment

This research was in part supported by “R&D promotion scheme funding international joint research” promoted by NICT, Japan.

References

Paul Buitelaar, Thierry Declerck, Nicoletta Calzolari, and Alessandro Lenci. 2003. Language Resources and the Semantic Web. In: *Proc. of ELSNET/ENABLER workshop*.

Nicoletta Calzolari. 2006. Community Culture in Language Resources - An International Perspective. In : *Proc. of LREC2006 Workshop Towards a Research Infrastructure for Language Resources*.

Lionel Clément, and Éric Villemonte de la Clergerie. 2005. MAF: a morphosyntactic annotation framework. In: *Proc. of LTC2005*.

Thierry Declerck. 2006. SynAF: Towards a Standard for Syntactic Annotation. In: *Proc. of LREC2006*, pp.229-233

Thierry Declerck, Nancy Ide, and Thorsten Trippel. 2008. Interoperable Language Resources. In: *Sprache und Datenverarbeitung (International Journal for Language Data Processing)*, to appear.

Scott Farrar, and Terry Langendoen. 2003. A Linguistic Ontology for the Semantic Web. *Glot International*, Vol.7, pp.97-100.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In: *Proc. of LREC2006*, pp.233-236.

Yoshihiko Hayashi. 2007. Conceptual Framework of an Upper Ontology for Describing Linguistic Services. In: Toru Ishida, Susan R. Fussell, Piek T. J. M. Vossen (Eds.): *Intercultural Collaboration*, LNCS 4568, Springer-Verlag, pp.31-45.

Ian Horrocks, et al. 2004. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. http://www.w3.org/Submission/SWRL/

Nancy Ide, and Laurent Romary. 2006. Representing Linguistic Corpora and Their Annotations. In: *Proc. of LREC2006*, pp.225-228.

Toru Ishida. 2006. Language Grid: An Infrastructure for Intercultural Collaboration. In: *Proc. of SAINT2006*, pp.96-100.

Toru Ishida, et al. 2008. A Non-Profit Operation Model for the Language Grid. In: *Proc. of ICGL2008*, to appear.

ISO DIS 24613:2007. 2007. Language resource management -Lexical markup framework (LMF), Rev.14.

Brigitte Jörg, and Hans Uszkoreit. 2005. The Ontology-based Architecture of LT World, a Comprehensive Web Information System for a Science and Technology Discipline. In: *Leitbild Informationskompetenz: Positionen - Praxis - Perspektiven im europäischen Wissensmarkt*.

Evan Klein, and Stephen Potter. 2004. An Ontology for NLP Services. In: *Proc. of LREC Workshop on a Registry of Linguistic Data Categories within an Integrated Language Resource Repository Area*.

Deborah L. McGuinness, and Frank van Harmelen. 2004. OWL Web Ontology Language Overview. http://www.w3.org/TR/owl-features/

Sue Ellen Wright. 2004. A Global Data Category Registry for Interoperable Language Resources. In: P*roc. of LREC2004*, pp.123-126.