What Does Interoperability Mean, Anyway?
Toward an Operational Definition of Interoperability for Language Technology

Nancy Ide
Department of Computer Science
Vassar College
ide@cs.vassar.edu

James Pustejovsky
Department of Computer Science
Brandeis University
pustejovsky@cs.brandeis.edu

Abstract

Two major projects in the U.S. and Europe have joined in a collaboration to work toward achieving interoperability among language resources. In the U.S., the project, Sustainable Interoperability for Language Technology (SILT) has been funded by the National Science Foundation under the INTEROP program, and in Europe, FLaReNet, Fostering Language Resources Network, has been funded by the European Commission under the eContentPlus framework. This international collaborative effort involves members of the language processing community and others working in related areas to build consensus regarding the sharing of data and technologies for language resources and applications, to work towards interoperability of existing data, and, where possible, to promote standards for annotation and resource building. This paper focuses on the results of a recent workshop whose goal was to arrive at operational definitions for interoperability over four thematic areas, including metadata for describing language resources, data categories and their semantics, resource publication requirements, and software sharing.

1 Introduction

Two major projects in the U.S. and Europe have joined in a collaboration to work toward achieving interoperability among language resources. In the U.S., the project, Sustainable Interoperability for Language Technology (SILT) has been funded by the National Science Foundation under the INTEROP program, and in Europe, FLaReNet, Fostering Language Resources Network, has been funded by the European Commission under the eContentPlus framework. This international collaborative effort involves members of the language processing community and others working in related areas to build consensus regarding the sharing of data and technologies for language resources and applications, to work towards interoperability of existing data, and, where possible, to promote standards for annotation and resource building.

A major condition for the take-off of the field of Language Resources and Language Technologies is the creation of a shared policy for the next years. FLaReNet aims at developing a common vision of the area and fostering a European strategy for consolidating the sector, thus enhancing competitiveness at EU level and worldwide. SILT's goal is to turn existing, fragmented technology and resources developed to support language processing technology into accessible, stable, and interoperable resources that can be readily reused across several fields.

This paper focuses on the results of a recent workshop whose overall goal was to arrive at an operational definition of interoperability. The workshop was motivated by the need to find ways to assess the current state of interoperability in the field of language technology as well as to measure our progress towards achieving interoperability in the future. An operational definition identifies one or more specific observable conditions or events and then tells the researcher how to measure that event; it must be valid (does it measure what they are supposed to measure?) and reliable (the results should be repeatable). The workshop considered interoperability over four thematic areas:

1. Metadata for describing language resources

2. Data categories and their semantics

3. Requirements for publication of data and an-

notations

4. Requirements for software sharing

For each of these areas, the concrete outcome of the workshop aimed to provide (1) an operational definition of interoperability; (2) an assessment of the current state of interoperability; (3) to the degree possible, consideration of how interoperability may be measured in that area; and (4) a "roadmap" of activities for the near and long term to bring us closer to the interoperability goal. In the remainder of this paper, we first discuss definitions of interoperability, and then summarize the discussions and conclusions of each of the working groups that addressed the four thematic areas.

2 Interoperability defined

Broadly speaking, interoperability can be defined as a measure of the degree to which diverse systems, organizations, and/or individuals are able to work together to achieve a common goal. For computer systems, interoperability is typically defined in terms of *syntactic interoperability* and *semantic interoperability*. Syntactic interoperability relies on specified data formats, communication protocols, and the like to ensure communication and data exchange. The systems involved can process the exchanged information, but there is no guarantee that the interpretation is the same. Semantic interoperability, on the other hand, exists when two systems have the ability to automatically interpret exchanged information meaningfully and accurately in order to produce useful results via deference to a common information exchange reference model. The content of the information exchange requests are unambiguously defined: what is sent is the same as what is understood.

For language resources, the focus is increasingly on semantic rather than syntactic interoperability. That is, the critical factor is seen to be the accurate and consistent interpretation of exchanged data rather than the ability to process it immediately without modification to its physical format. The reasons for this are several, but first and foremost is the existence of large amounts of legacy data in varied syntactic formats, coupled with the continued production of resources representing linguistic information in varied, but mappable, ways. Indeed, to ensure interoperability for language resources, the trend in the field is to specify an *abstract data model* for structuring linguistic data to which syntactic realizations can be mapped, together with a mapping to a set of *linguistic data categories* that communicate the information (linguistic) content. In the context of language resources, then, we can define syntactic interoperability as the ability of different systems to process (read) exchanged data either directly or via trivial conversion. Semantic interoperability for language resources is virtually the same as for software systems: it can be defined as the ability of systems to interpret exchanged linguistic information in meaningful and consistent ways reference to a common set of reference categories.

3 Metadata for describing language resources[1]

3.1 Current situation

Metadata for language resources provides information that enables users to identify and locate language resources that exhibit a set of characteristics matching the user's needs. Syntactic interoperability among metadata specifications for language resources is ensured via a common set of *metadata labels*; semantic interoperability is ensured if these labels correspond to identically defined *metadata categories*. Both syntactic and semantic interoperability for language resource metadata have received some attention over the past several years through efforts such as the Open Language Archives Community (OLAC)[2]. However, language researchers must still master multiple metadata sets in order to search multiple locations in order to find all the description available for needed resources or else risk failing to note the existence of critical LRs and then either recreate them or else do without them.

Currently, major data centers (e.g. ELRA, LDC) maintain their own separate catalogs using different metadata languages (categories, terminologies) and export subsets of their metadata categories to the OLAC central data repository. OLAC provides specifications for OAI (Open Archives Initiative[3]) compliant metadata as well as routines for harvesting, interchanging and searching their metadata. ELRAs universal catalog seeks to extend the work of OLAC while

[1] Working group members were Nicoletta Calzolari, Khalid Choukri (co-leader), Christopher Cieri (co-leader), Terry Langendoen, Johannes Leveling, Martha Palmer, and James Pustejovsky.

[2] http://www.language-archives.org

[3] http://www.openarchives.org

focusing on resources intended for HLT R&D. Specifically, the ELRA UC includes a greater percentage of ELRA metadata fields and exploits data mining to discover resources not produced or distributed by ELRA. The LREC Map is an initiative to exploit the biennial LREC (Language Resources and Evaluation Conference[4]) abstract submission process as a way to increase the contribution of LR metadata. The NICT Shachi[5] catalog also attempts to serve as a union catalog of resources including resource produced by NICT and elsewhere. Shachi currently differs from OLAC in that catalog records are scraped rather than harvested as part of a bilateral negotiation. Shachi also uses data mining technologies to discover information about LRs that may not be present in their home catalog entries. The LDC LR wiki[6] identifies LRs for less commonly taught languages organized by language and LR type with individual sections edited by area experts. The wiki permits free text description and intends to attempt normalization as an activity under the current proposal. Finally, the LDC LR Papers Catalog[7] enumerates research papers that introduce, describe, discuss, extend or rely upon another LR. A summary of the current situation is given in Table 1.

3.2 Definition of Interoperability for Metadata

We define interoperability of metadata languages L_1 and L_2 as the capability of two metadata providers to interchange metadata records m_1 written in L_1 and m_2 written in L_2 for a single LR r via a function f that maps L_1 to L_2, such that a query that returns r in L_2 also returns r in $f(L_1)$. Less formally, a single search should work equally well, retrieving the same LRs, when issued against different but interoperable catalogs assuming the same base metadata in the two catalogs perhaps mapped to a new form in one of the catalogs.

Given this definition, it is clear that the situation today as described above does not fulfill the requirements for interoperability for metadata. To address this problem, we propose a Roadmap of activities consisting of two efforts, one for the short term and another for the long term.

3.3 Roadmap

For the short term, the working group proposed a simple and concrete effort to harmonize the LR catalogs of the largest international data centers, including initially those of ELRA and LDC as well as the Shachi union catalog produced by NICT. This will not be an attempt to identify the minimal subset of the metadata fields that apply to all LR types; rather, it will begin with focus on the domain of LRs targeted toward HLT R&D and identify the superset of metadata types contained in them. The next step will be to review the types, one by one, to identify those than can be normalized internally and across data centers and distinguish those from the smaller subset that encode irreconcilable differences among the business practices of the centers as adapted to their local regulatory constraints. The data center partners will agree to normalize and harmonize practice wherever possible. The outcome of this part of the effort will be new fully functioning LR catalogs to replace those of the data centers partners as well as a definition of the metadata categories, a database structure to hold the metadata and a search engine customized to HLT LR search.

These new resources along with a specification of best metadata practices will be made available to other data centers and individual data creators to use in the creation of their own catalogs. To promote the sustainability of LR held outside data centers, a centralized metadata repository with a harvesting protocol will be provided. To address the range of competences among LR searchers, the search engine will permit both use of controlled vocabulary fields and relevance based search of entire catalog records. A novel contribution of this effort will be searcher assistance based upon the relations among metadata categories (*dictionary* $\simeq$ *lexicon*) and prior search behavior (those who searched for *Gigaword* also searched for *news text corpora*). Coupled with searcher assistance, we will provide metadata creator assistance based on searcher behavior and behavior of other metadata providers ("93% of searchers include a language name in their search"but "87% of all providers include ISO 639-3 language codes" and "the metadata you have provided so far also characterize 32 other resources"). In order to effectively manage the harmonization of data center catalogs and the provision of metadata resource, we will construct a governance body specifically for this

[4] http://www.lrec-conf.org

[5] http://www2.shachi.org

[6] http://lrwiki.ldc.upenn.edu

[7] The LDC LR Papers Catalog is currently a local effort undertaken with LDC discretionary funds. Once the Catalog has reached an appreciable size, it will be opened to the community and additions from remote authors will be accepted.

	OLAC	ELRA UC	LREC Map	NICT Shachi	LDC LR Wiki	LDC Papers Catalog
external resources	X	X	X	X	X	X
normalized metadata	X	X	X		deferred	X
raw resources					X	
scraping				X	X	
data mining		X		X		
papers as LRs	X		X			X

Table 1: Summary of Metadata Efforts

project. The group with include representatives of the project partners, sponsors, individual and small group resource providers and of LR users.

For the longer term, the working group proposed to expand the scope of the universal catalog to include two important and frequently overlooked LRs on either end of the processing spectrum: raw unprocessed data and the most carefully processed LRs, research papers. Some of this work has already begun in a number of individual efforts that have not been coordinated across this same span of data centers and LR creators. Specifically, a Less Commonly Taught Resource (LCTL) Language Resource wiki was developed by LDC within the REFLEX program. Similar efforts to harvest papers describing LRs are underway at LDC using human effort and within the Rexa project[8] using data mining technologies. Our proposal here is to accomplish this expansion by data type while integrating model workflow methodologies into the workflow including social networking, web sourcing, and data mining. In this project, our intent is to enhance the universal catalog with links to raw resources including web sites rich in monolingual and parallel text and lexicon built for interactive use. The resources are necessary to advance the universal catalog toward the very apt goal of true universality as it affects languages whose representation among formal LRs is insufficient with respect to their global importance. Those who would create HLTs for these languages must resort to primary LR resource creation based upon harvests of these raw resources. As the project moves from short to middle terms objectives it will be necessary to adjust its governance and broaden the scope of its normalization activities. The issues that challenge this expansion to raw resources and papers differed markedly from those that challenge the harmonization of traditional catalog metadata and the collaborators must change in response. The principles centers of these middle term activities will be large data providers that have already implemented sustainable business models.

4 Data Categories and their semantics[9]

4.1 Current situation

Semantic interoperability as defined in Section 2 is the most critical for language resource interoperability because the information shared among systems relies critically upon a shared definition of linguistic elements. Semantic interoperability at this level is also the most difficult to achieve.

Efforts in the 1990's were devoted to establishing standard sets of data categories, most notably within the European EAGLES/ISLE project[10], which developed standards for morpho-syntax, syntax, sub-categorization, text typologies, and others. However, none of these standards has achieved universal acceptance and use.

The most recent large-scale effort addressing standardization of data categories (among other topics) is ISO TC37 SC4 (Language Resource Management), which has established a prototype of a *data category registry* (DCR)[11] containing many low-level linguistic categories and their definitions. The ISO approach is to provide a set of data categories defined by experts in the field. Each data category is assigned a unique identifier together with a linguistic descriptions consisting of a definition, specification of associated value domains, and examples. Via reference to its identifier, a data category or category can be associated with any data element name used in a language resource and/or language-specific versions of definitions, names, value domains and other attributes. Because the DCR categories include very granular elements, they can serve as building blocks for composing more complex linguistic descriptors. As such, the data categories in the registry

[8] http://rexa.info

[9] Working group members were Dan Flickinger, Jerry Hobbs, Nancy Ide, Monica Monachini, Jan Ojdik (leader), and Nianwen Xue

[10] http://www.ilc.cnr.it/EAGLES96/browse.html

[11] http://www.isocat.org

provide the “common information exchange reference model” required for semantic interoperability.

The working group

risks of going for a full interlingual approach But the only way to find out is pursue this

A less ambitious approach with (multiple sub-

Fundamental Questions

5 Publication of Resources[12]

5.1 Motivation

Currently, no guidelines or even common practices exist for creating, documenting, and evaluating language resources, including text and speech corpora, linguistic annotations, lexicons, grammars, and ontologies, that are “published”–i.e., made available for use by others. Some standard practices for resource publication through established data distribution centers such as LDC or ELRA exist, but even these are not completely consistent among different centers, and they are not comprehensive. More crucially, many resources are made available via web distribution, and the format of the resource and information about creation methodology and resource quality is highly variable and in some cases, non-existent. Given the recent increase in resource production, the need for standardized procedures for publishing resources is rising. Users need information to assess the quality of a resource, to replicate processes and results, and to deal with idiosyncrasies or documented errors. This kind of documentation is very often unavailable or difficult to acquire.

Clear guidelines for resource publication will impact the resource creation process, by specifying requirements for quality assurance and implicitly establishing baselines for cost, time frame, and requisite facilities, all of which are for the most part unknown at this time. Furthermore, such guidelines will inform standard procedures for resource evaluation, by establishing both clear specifications for documenting a published resource that will figure into the evaluation itself and self-evaluation metrics that should accompany the published resource. Therefore, a set of standards for resource publication bears on several aspects of interoperability, some of which were addressed by other working groups at the Brandeis meeting.

Due to the lack of established procedures and practices, the fundamental question addressed by this group was therefore “What set of requirements for the release or publication of a data resource maximizes the potential usefulness and interoperability of that resource?” To answer this question, the working group identified two broad types of requirements: (1) formats and access, and (2) documentation (taken in the broadest sense). Because (1) has received considerable attention by other groups and efforts and some best practices are already established, specifications for resource documentation proved to most in need of consideration in order to ensure that published resources are immediately usable and interoperable. Each of these requirements is addressed in the sections that follow.

5.2 Formats and access

5.2.1 Standardized formats

All resources should be released in a *standard format*, as defined in Section 2, that is, a format that provides for both syntactic and semantic interoperability. Semantic interoperability is the topic of Section 4 and will not be addressed here.

Several *de facto* best practices for language data annotation, in particular, are emerging that we can espouse here as characteristics of interoperable resources. They include:

Independence of source. Primary *language* data[13] are *read-only*, and annotations are in *stand-off* form referencing the primary data or other annotations.

Conformance to a Common Representation Model. Over the past decade, the directed graph has emerged as the *de facto* standard abstract model for linguistic annotations.[14] There are several issues here. First is the question of mappability to the abstract model; in general, all annotations are mappable to the directed graph. Second is the question of mappability to a concrete serialization of the model, such as ISO/GrAF (Ide and Suderman, 2007) or UIMA CAS. The final is-

[12] Working group members were Nancy Ide, Aravind Joshi, James Pustejovsky, Ineke Schuurman, Satoshi Sekine, Claudia Soria, and Marc Verhagen (leader).

[13] If the original content is for example in HTML or Microsoft Word, there may be a pre-processing phase that extracts the language data from the sources. Similarly, if data need to be anonymized, the primary read-only language data are the anonymized versions of the sources.

[14] Note that the graph model underlies recent widely-used formats such as Annotation Graphs (Bird and Liberman, 2001), XML and RDF models, UIMA CAS, and is formalized in the ISO Linguistic Annotation Framework (Ide and Suderman, 2006).

sue is whether a mapping is actually provided, and at what level (e.g., a declarative specification or a program that generates the mapped version).[15]

Encoding. Best practices for language data are converging on the use of UTF-8 for primary data in western languages.

5.2.2 User Access

Registration of the resource. A published resource should be listed in resource catalogs like ELRA, LREC Map, or LDC, even if this is not the primary distribution source. The resource creators should provide the information required by those catalogs in order to be listed appropriately.

Sustainability. Means for resource preservation and maintenance should be established prior to publication to ensure continued availability. One means to ensure sustainability is to distribute the resource solely through an established resource center such as LDC or ELRA. In the case where resources are distributed via the web (e.g., a website local to the organization that developed the resource, or a web distribution mechanism such as Sourceforge or CPAN), ensured sustainability is the responsibility of the resource developer.

Metadata. A published resource should be accompanied by appropriate metadata relevant to the resource type. The metadata recommendation should as much as possible be based on existing metadata specifications. Metadata should include (1) formal specifications meant for machine processing; (2) metadata for the resource as a whole (e.g., corpus, lexicon, etc.); and, where applicable, metadata for each resource component (e.g. individual texts or text collections in a corpus (name, source, author, etc.; individual annotation types (e.g. tokenization, named entities, co-reference). Section 3 provides additional information on metadata requirements for interoperability. However, metadata requirements for resources such as annotations are only now being established in the ISO LAF specification (ISO TC37 SC4 WD24611); this is an area that needs attention.

5.3 Documentation

We recognize several different kinds of documentation, which may exist in one or several physical documents or in header(s) associated with data and annotations. Each is designed to meet the needs of certain users of the resource. All document types may not be applicable to all resources. Also, some documentation types overlap with metadata as covered in Section 3. The documentation types identified are:

i. *High-level description*: provides the non-expert, interested reader a good idea of what is in the resource.

ii. *Annotation/resource creation guidelines*: guidelines directly used by annotators, validators, or creators of the resource (e.g. creators of lexicon or ontology entries, etc.). These may be in the local language. A more global version in English should be provided when possible.

iii. *Background*: information on the theoretical framework, background, and/or the “philosophy” of the resource.

iv. *Methodology*: a precise specification of the methodology used to create the resource. This information should be specific enough to enable others to replicate the process and obtain the same results. It should include

- full documentation of tools used in any phase of the creation process, including software with version, source of the software, software documentation or publication, and an indication of the platform the tools were run on.
- data preparation methods, including information about the data source, normalizations/corrections performed, etc.
- error rates and manual validation/corrections for automatically produced annotations;
- description of annotation and quality control procedures, including standard inter-annotator agreement statistics (Kappa, P&R, TBD) if more than one annotator annotated the same document.

v. *Description of category semantics*: prose specification of the data categories and their

[15] Mappability to the common annotation model may be relevant to resource types other than annotations, since formats such as ISO LAF generalize over standard formats for other resource types (e.g., ISO Lexical Markup Format).

semantics, with substantial examples from the resource. This should include documentation of the evolution of specifications (if versioned), illustrating the learning process.

vi. *Formal specifications*: XML-Schema, RDF schema, formal metadata specifications, grammar for annotation syntax, etc.

vii. *Project documentation*: Project description, location, personnel, contact. Statistics: funding source, costs in person hours to create the resource.

viii. *Data documentation*: Corpus information: source, original format, errors in the data, trustability of source, OCR error rate (if applicable), copyright notice for data documents included in the resource (if different from copyright for entire corpus), and a specification of which annotations may apply. Speech resources: how was the signal required, participants, etc. Much of the information needed here can be found in the TEI Header.

ix. *Resource documentation*: Release date, version history, usage restrictions/copyright notice, availability, LDC or ELRA catalog number.

x. *Supporting materials*: tutorials, presentations, published papers.

5.4 Roadmap

For the short term, we propose several steps to solidify the information required to accomplish the long-term goal of providing a full specification of the requirements for resource documentation that support interoperability. First, it is necessary to involve representatives of the speech and multimedia communities to ensure that requirements for resources of those types are accommodated. Second, we propose a review of literature on methodology for language resource creation, insofar as it exists. Potential sources include ELRA's specifications for production, validation, distribution, and maintenance of language resources; LDC's data creation methods[16], reports from earlier projects such as EAGLES and Eurotra, and "The Production of Speech Corpora" Cookbook[17]. For other media, a list of *de facto* or best practice standards must be compiled. Third, we need to address metadata for second-order resources such as annotations, which are only now beginning to be addressed in the ISO LAF specification (ISO TC37 SC4 WD24611). We propose to review the ISO recommendations for completeness and make suggestions for modifications or additions if appropriate.

Conformance to a common model (see Section 5.2.1)

6 Software Sharing[18]

With respect to software, interoperability describes the capability of different language resources (X) to operate jointly via a common set of data categories (W), to read and write the same file formats (Z), and to use the same protocols (Z). The goal of software sharing is to make language tools and resources used for scholarly research available to other investigators and developers across diverse communities and compatible with one another: Academia, Research institutions, Government, and Industry. To this end, three areas of interoperability are relevant to our concerns:

- Software formats
- Data formats
- Software integration platforms

The ability to combine different tools has been greatly improved by software integration platforms such as GATE and UIMA. Many vendors/providers are now offering UIMA-ready components; GATE and UIMA each have interoperability layers; and there are integration components with (meta)annotation frameworks such as GrAF. This has resulted in interoperable annotation structures, thereby providing an integrated solution for both data and tools (e.g. UIMA CASs). This is a starting point for linguistically annotated resources, but it does not provide for other resource types such as lexicons and ontologies. Furthermore, adoption of the UIMA CAS model–or anything similar–is far from universal, most importantly because significant "legacy" resource creation projects, which serve as default

[16] http://www.ldc.upenn.edu/Creating/

[17] http://www.phonetik.uni-muenchen.de/forschung/BITS/TP1/Cookbook/

[18] Working group members were Ed Loper, German Rigau, and Antonio Sanfilippo (leader)

standards/models for projects that extend the resource to other languages, do not adhere to this model.

Evaluation is an important component to ensure interoperability for software sharing, including comparison of different approaches to a given problem and the ensuing applications, assessment of the availability of resources and technologies for a given application, assessment of the state-of-the-art for a given technology, and benchmarking system usability and user satisfaction.

Finally, some of the challenges facing the issue of software interoperability are as follows:

- Software engineering:
  - What is the right balance between code abstraction and code implementation?
- Data formats:
  - How do we deal with standards proliferation?
  - How do we balance across operational and research needs?
- Software integration platforms:
  - Can annotation structure function as the unifying factor?
- How do different licensing agreements influence access and distribution?
- How do patenting and copyright affect access and distribution?
- How do we reconcile differences across data and software sharing policies?
- How do we develop proper metrics to evaluate the utility of language data and tools?
- How do we address privacy and security?

6.1 Roadmap

The first item on the roadmap can be identified as *making LRs more discoverable*. The second item can be identified as *making legacy and new LRs interoperable*:

- Define a specification language that describes existing LR formats, taking into account existing LE standards
  - Needs to be able to describe a very wide variety of formats, e.g. one sentence per line text document, parenthesized trees w/ standard treebank tags, ASCII raw text, TimeML, PAROLE, Penn Treebank, FrameNet, VerbNet, PropBank,
- Identify best practices to guide development of new LRs to ensure interoperability Includes actions on formats, licensing, IPR and policies, and roles and responsibilities

The final roadmap item can be identified as *making LRs amenable to evaluation*:

- Define metrics for each language resource
- Define tests for assessing metrics:
  - Measure accuracy, precision/recall, etc.
  - Measure utility and usability: implement "social intelligence approaches, such as Amazon/eBay type recommendation systems and make available to all stakeholders.

Acknowledgments

This work was supported by U.S. National Science Foundation grant INT-0753069.